NASA/TP–2003–210791

Cox Proportional Hazards Models for Modeling the Time To Onset of Decompression Sickness in Hypobaric Environments

Laura A. Thompson, Ph.D., M.A
University of Houston – Clear Lake
School of Natural and Applied Sciences

Raj S. Chhikara, Ph.D., M.S.
University of Houston – Clear Lake
School of Natural and Applied Sciences

Johnny Conkin, Ph.D., M.S.
NASA/Johnson Space Center

March 2003

THE NASA STI PROGRAM OFFICE . . . IN PROFILE

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA's counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or cosponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and mission, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at http://www.sti.nasa.gov

- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA Access Help Desk at (301) 621-0134

- Telephone the NASA Access Help Desk at (301) 621-0390

- Write to:
  NASA Access Help Desk
  NASA Center for AeroSpace Information
  7121 Standard
  Hanover, MD 21076-1320

NASA/TP–2003–210791

Cox Proportional Hazards Models for Modeling the Time To Onset of Decompression Sickness in Hypobaric Environments

Laura A. Thompson, Ph.D., M.A
University of Houston – Clear Lake
School of Natural and Applied Sciences

Raj S. Chhikara, Ph.D., M.S.
University of Houston – Clear Lake
School of Natural and Applied Sciences

Johnny Conkin, Ph.D., M.S.
National Space Biomedical Research Institute

March 2003

Contents

Tables

Figures

Acronyms and Nomenclature

AFT	accelerated failure time
AIC	Aikaike's Information Criterion
API	adjusted prognostic index
CDF	cumulative distribution function
DCS	decompression sickness
D.P2	The deviation, $P2 - \overline{P2}$, where $\overline{P2}$ is the arithmetic average of the values on the covariate, P2
EVA	extravehicular activity
EXER	indicator variable of whether exercise is done repetitively during exposures (EXER = 1) or not (EXER = 0)
F-H	Fleming-Harrington
HDSD	hypobaric decompression sickness databank
ISS	International Space Station
ISSO	Institute for Space Systems Operations
JSC	Johnson Space Center
KM	Kaplan-Meier
LOWESS	locally weighted least squares estimation (typically used to fit a smooth curve through a scatter plot of points)
LogLH	log likelihood
LRT	likelihood ratio test
MPLE	maximum partial likelihood estimate
P2	ambient pressure (in psia) at the final altitude
PH	proportional hazards
PI	prognostic index
psia	pounds per square inch absolute
PN2360	partial pressure of nitrogen (in psia) in the 360-minute half-time compartment
TR360	= (PN2360/P2), tissue ration in the 360-minute half-time compartment
VGE	venous gas emboli

Acknowledgments

The Institute for Space Systems Operations of the University of Houston provided basic funding support for Dr. Laura Thompson under the Institute for Space System Operations Post-Doctoral Fellowship Program. Partial support of her research was under research grant NASA 9-1083 from the NASA/Johnson Space Center, Houston. The authors thank Dr. Alan Feiveson for his helpful comments and suggestions on an earlier draft of the report.

Abstract

In this paper, we fit Cox proportional hazards models to a subset of data from the Hypobaric Decompression Sickness Databank (Conkin et al., 1992). The data bank contains records, accumulated from literature sources and experiments, on the time to decompression sickness (DCS) and venous gas emboli for over 130,000 person-exposures to high altitude in chamber tests. The subset we used contains 1,321 records, with 87% censoring, and has the most recent experimental tests on DCS that have been made available from Johnson Space Center. We built on previous analyses of this data set by considering more expanded models as well as more detailed model assessments specific to the Cox model. We found that a Cox model stratified on the quartiles of the final ambient pressure at altitude, which is one of the covariates, described the data better than did previously considered models (e.g., English 2000, Chhikara et al., 1998). Our model also included final ambient pressure at altitude as a nonlinear continuous predictor, as well as the computed tissue partial pressure of nitrogen at altitude, and whether or not exercise was done at altitude. We conducted various assessments of our model, many of which were recently developed in the statistical literature, and we concluded where the model needs improvement. We considered the addition of frailties to the stratified Cox model, but found that no significant gain is attained over a model that does not include frailties. Finally, we validated some of the models we fit. The results in this paper serve as a useful addition to the growing literature on statistical analysis of DCS in hypobaric environments.

1. Introduction

When humans travel to a hypobaric environment, gas that normally dissolves in tissues can escape from solution to form gas spaces or bubbles that can displace or damage tissues (Conkin, 2001). The displacement of tissues by trapped gas spaces can cause a wide variety of symptoms, ranging from the mild joint pain in the elbows, knees, and shoulders that is commonly called the 'bends' (Type I decompression sickness (DCS)) to more serious symptoms that result from bubbles lodging near major organs such as the brain (Type II DCS). The presence of these symptoms is collectively referred to as DCS. In general, the longer the exposure to pressure reduction, the greater the risk of DCS, especially of Type II DCS.

Although the space shuttle or the airlock on the International Space Station (ISS) is pressurized to sea-level pressure (14.7 pounds per square inch absolute (psia)), crew members are exposed to a greatly reduced absolute pressure during extravehicular activity (EVA) because their spacesuits are pressurized to much lower than sea level (about 4.3 psia) (Foster et al., 1998). So astronauts risk Type II DCS each time they perform EVAs. The risk can be decreased or prevented through sufficient denitrogenation prior to EVA. Denitrogenation procedures typically involve prebreathing pure oxygen or oxygen-enriched mixtures. One of the goals of this paper is to build a model that quantifies the risk of DCS in hypobaric situations similar to those experienced by astronauts on EVAs.

Several researchers have fit parametric survival models to right-censored DCS data from altitude exposures. Kumar and Powell (1994) modeled log time to onset of DCS as a parametric function of explanatory variables such as tissue ratio (a measure of nitrogen decompression stress) and the presence of circulating microbubbles in venous blood. Kannan and Raychaudhuri (1998) and Conkin et al. (1996) fit log-logistic models to time to onset of DCS symptoms. Together these studies included covariates such as tissue ratio, type of exercise at altitude, final pressure at altitude, and prebreathing time. In addition, Kannan and Raychauduri (1998) fit a semi-parametric Cox (1972) proportional hazards model and concluded its predictions were very close to those of the log-logistic model. More recently, Conkin (2001) took an evidence-based approach to estimate a nonlinear parametric hazard function, and then used the integrated hazard as a component of the probability in a Bernoulli likelihood predicting whether or not Type II DCS occurred in a series of published studies. Later, Thompson and Chhikara (2001) fit a random effects logistic model to the same data.

The objective of this study is to model onset time to DCS and to assess the effect of certain covariates. To do this, we analyzed a subset of the Hypobaric Decompression Sickness Databank (HDSD) (Conkin et al., 1992) that was accumulated from literature sources and experiments at Johnson Space Center (JSC). The HDSD contains records from over 130,000 person-exposures to high altitude in chamber tests. The subset we used contains 1,321 records from the most recent experimental tests on DCS made available from JSC to develop safe decompression procedures

for EVA. The general model we used for this data set is the Cox (1972) model because of its attractive semi-parametric nature and ease of fitting.

This subset of the HDSD has been analyzed previously. Chhikara et al. (1998) initially examined accelerated failure time models for these data (lognormal and log-logistic) but preferred the Cox proportional hazards model because it does not assume a parametric distribution for the failure time. These authors included as covariates the computed tissue partial pressure of nitrogen at altitude, ambient pressure at altitude, and whether exercise was done at altitude. English (2000) also analyzed these data and included an interaction term and fit models that was stratified on whether or not exercise was done at altitude. The set of models we considered differs from previous models in that we included a different set of covariates and considered stratification on other variables. Furthermore, we included an extensive set of diagnostics to evaluate model fits. From this we show the models considered previously for this data set are inadequate because they do not account for all of the variability in the data.

In Section 2, we give a brief introduction to Cox's proportional hazards model and its extension to stratification. In Section 3, we describe the data and the covariates we included in the models. We examine estimated survival functions and hazard functions by covariates. In Section 4, we discuss the assessment of the proportional hazards assumption in the Cox model. In Section 5, we examine fitting stratified Cox models to the DCS data and conduct residual analysis. In Section 6, we describe one of the final chosen models for these data. In Section 7, we discuss fitting frailty models to the data. Finally in Section 8, we validate some of the models we have considered.

2. Cox Proportional Hazards Model

In survival models, the hazard function for a given individual describes the instantaneous risk of experiencing an event of interest within an infinitesimal interval of time, given that the individual has not yet experienced that event. Cox (1972) proposed a semi-parametric model for the hazard function that allows the addition of explanatory variables, or covariates, but keeps the baseline hazard as an arbitrary, unspecified, nonnegative functional of time. The Cox hazard function for fixed-time covariates, **x**, is

$$l(t;\,\mathbf{x}) = l_0(t)\exp(\mathbf{x}'b) \tag{2.1}$$

Due to the construction of (2.1), the baseline hazard $l_0(t)$ is defined as the hazard function for that individual with zero on all covariates. Because the baseline hazard is not assumed to be of a parametric form, Cox's model is referred to as a semi-parametric model for the hazard function. The survival function corresponding to (2.1) is then (e.g., Lawless, 1982)

$$S(t;\mathbf{x}) = \exp\left[-\exp(\mathbf{x}'b)\int_0^t l_0(u)du\right] \tag{2.2}$$

The integral in (2.2) is called the baseline cumulative hazard function. Several methods are available for estimating the baseline cumulative hazard function (Klein and Moeschberger, 1997).

Cox's model has become the most used procedure for modeling the relationship of covariates to a survival or other censored outcome (Therneau and Grambsch, 2000). Its form is flexible enough to allow time-dependent covariates as well as frailty terms and stratification. However, it has some restrictions. One of the restrictions to using the Cox model with time-fixed covariates is its proportional hazards (PH) assumption; that is, that the hazard ratio between two sets of covariates is constant over time. This is due to the common baseline hazard function canceling out in the ratio of the two hazards. Thus, for fixed-time covariates, the exponent of a coefficient describes the relative change in the baseline hazard due to that covariate.

The baseline hazard is typically considered a 'nuisance parameter,' and estimation of b is done by maximizing a profile likelihood with $l_0(t)$ being substituted for an expression involving b and $\mathbf{x}$, as well as the times at which

failures occurred (Klein and Moeschberger, 1997). This expression is called the profile maximum likelihood estimate of $l_0(t)$. The likelihood with $l_0(t)$ 'profiled out' is called the *partial likelihood* by Cox (1972). For fixed-time covariates and independent observations, the partial likelihood is

$$L(b) = \prod_{i=1}^{D} \frac{\exp(\mathbf{x}_i' b)}{\left[\sum_{j \in R(t_i)} \exp(\mathbf{x}_j' b)\right]^{d_i}} \tag{2.3}$$

where *D* is the number of events, d_i is the number of events at time t_i, and $R(t_i)$ is the risk set at time t_i (the number of subjects in the data set who have recorded DCS times or censored times later than or equal to time t_i). The value of b that maximizes (2.3) is called the maximum partial likelihood estimate (MPLE).

2.1. Stratified Cox Models

The model in (2.1) can be extended to account for stratification. The strata divide the subjects into disjoint groups, each of which has a distinct (arbitrary) baseline hazard function but common values for the coefficients b (Therneau and Grambsch, 2000). The hazard function for an individual *i* who belongs to stratum *k* is then

$$l(t; \mathbf{x}_i) = l_k(t) \exp(\mathbf{x}_i' b) \tag{2.4}$$

Typically, strata are naturally defined within the context of the problem. For example, multi-center clinical trials typically stratify on the clinic in which they are conducted (Therneau and Grambsch, 2000). However, the stratified Cox model also allows a deviation from proportional hazards, and as such provides an alternative to the assumption of proportional hazards. The hazard functions for two different strata do not have to be proportional to one another. However, within a stratum, proportional hazards are assumed to hold. We take advantage of this use of stratification for the DCS data.

The partial likelihood for stratified Cox models with *K* strata becomes a product of *K* terms, each of the form of (2.3), but where *i* ranges over only the subjects in stratum *k* ($k = 1,\ldots,K$).

3. Description of Data

The HDSD (Conkin et al., 1992) contains records for groups of human volunteer subjects who were exposed together to high altitude in one chamber test. Exposure records used for this analysis are from 1,321 individuals who participated in chamber tests at JSC. This subset of the data was originally extracted for use with a different analysis (Conkin, et al., 1998). The criterion for selection of a record in that study was that the record contained certain detailed information about venous gas emboli (VGE)—the movement of gas bubbles into venous blood—whose presence was proposed to be a precursor of DCS symptoms. The records also contained information on the reported onset of DCS symptoms. One thousand three hundred and twenty-two records were selected based on this criterion. In a previous analysis of DCS data (Chhikara et al., 1998), one observation had been discarded due to its high value on a measure of influence, leaving 1,321 records. These are the records that we use for this analysis. In this paper, we do not consider the information on VGE in the data set for any analysis here.

Although subjects were tested in groups, typically of around three individuals each, the grouping information was not recorded in the data set. Each test involved one decompression. During a test, subjects in a group were monitored for Doppler-detectable gas bubbles, and the test was terminated for a subject either upon reported incidence of a DCS symptom or, if the subject did not experience DCS, when the test period was over. Total test times at altitude ranged from 20 minutes to over 12 hours, with a mean of 4.66 hours. Onset of DCS was recorded for a subject if that subject reported any sign or symptom of Type I DCS. Thus, an observation for a subject was either

his or her reported DCS onset time or the test termination time, whichever came first. If it was the test termination time, the observation was considered Type I right-censored (Lawless, 1982). The data set contained 1,154 right-censored observations, or a little over 87% of the total records.

For some tests, subjects were assessed at hourly intervals for DCS symptoms to ensure they were not neglecting to report symptoms that were present. Because there were some individuals whose reported DCS onset was at one of the hourly checkpoints, it may be that the symptoms of DCS occurred as far back as almost one hour prior to what was actually recorded. The subjects for whom this was true were not indicated in the data set, but 25 records had reported DCS times on the hour. We treated these DCS times as 'exact' because the way in which they were reported may reflect how some DCS symptoms are reported during an EVA. Because of the importance of the task for which an EVA is required, the time at which DCS symptoms are reported by a spaceflight crew member may not be when the symptoms actually begin, but may be reported later after some inquiry.

In addition, a number of explanatory variables were measured for each group. In our analysis, we considered the explanatory variables listed in Table 1.

Table 1: Measured Explanatory Variables

	EXER	P2 (psia)	PN2360 (psia)	TR360
Minimum	0	3.000	4.032	0.938
Mean	0.801	6.203	9.198	1.536
Median	1	6.000	10.245	1.454
Maximum	1	10.110	12.320	3.453
SD	0.399	1.953	2.326	0.342

The first variable in Table 1, EXER, is an indicator variable showing whether the subject exercised repetitively during his or her time at altitude. Repetitive exercise is done during the tests to simulate vigorous work activity required on an actual EVA. P2 is the final environmental pressure reached in the exposure in psia. PN2360 is the calculated partial pressure of nitrogen (in psia) in a designated theoretical tissue compartment after prebreathing any oxygen-enriched mixture prior to ascent in the chamber. The calculation is obtained from a nitrogen-elimination model detailed in Conkin et al. (1996). The model states that the partial nitrogen pressure reached in the compartment after a specific time is a function of the initial nitrogen partial pressure in the tissue compartment at sea level and of the ambient nitrogen partial pressure in the prebreathe mixture. A theoretical compartment with a half-time elimination of 360 minutes was used in the model. Half-time is the time it takes to decrease to one-half of the difference in the initial nitrogen pressure minus final nitrogen pressure. The 360-minute half-time compartment was chosen in a different study out of a spectrum of different half-times ranging from 300 to 540 minutes. Each of these half-time compartments was used in a model fitted to a data set obtained from the HDSD (Conkin et al., 1996). The 360-minute half-time was chosen in that study based on log likelihood calculations and relevance to JSC.

The last variable, TR360, was not measured itself, but is a ratio of the preceding variables, PN2360 to P2. The ratio represents a decompression stress index. When this ratio exceeds the fraction of nitrogen pressure present at sea level (11.6/14.7 = 0.78), we would expect DCS to occur more quickly. In the models we consider, we do not need to model both PN2360 and TR360 as covariates, and only use TR360. TR360 is more easily interpreted as a unitless index of decompression stress, and can be used to make direct comparisons. Although PN2360 (and thus TR360) is actually computed from an assumed model that may not be completely accurate, it is nonetheless treated here as though it is measured without error. However, the calculated TR360 values are frequently used directly by NASA in their tests of decompression procedures. Conkin et al. (1996) contains further information on the origination of the TR360 index.

Subjects were all determined to be in good physical condition and received the necessary orientation to the chamber prior to taking part in the tests. The average ages of both males and females was 31 years. Of the 1,321 subjects, 1,030 were males and 291 were females. However, sex was not used as a covariate in this analysis. In the types of decompression tests that are represented in the data set, gender is frequently not an important predictor of

DCS. Indeed, preliminary analysis suggests that sex does not add predictability to the models we consider here. A more detailed description of the subjects in this analysis can be found in Conkin et al. (1998).

3.1. Exploratory Analysis

To provide some initial insight into the characteristics of the data and to facilitate further discussion, we constructed several cross-tabulations of explanatory variables by proportion of DCS cases. For the tables, continuous variables were categorized into quartiles. Table 1a shows the DCS proportions for TR360 and EXER, and Table 1b shows the proportions for P2 and EXER. Table 1a also shows that those subjects with higher TR360 had increasingly greater incidence of DCS, and this trend did not change much across EXER. The marginal totals across TR360 categories show that those subjects who exercised at altitude had a slightly greater incidence of DCS symptoms. Table 1b shows that in general, subjects exposed to a lower final ambient pressure had a greater incidence of DCS. But, the highest incidence category was the (4.30, 6.00] category.

Table 1a: Proportion DCS by TR360 and EXER

TR360 categories				
(0.93, 1.31]	**(1.31, 1.45]**	**(1.45, 1.78]**	**(1.78, 3.45]**	**Marginal Proportions (EXER)**
EXER = 0				
2/64 = 0.03	1/91 = 0.01	11/63 = 0.17	13/45 = 0.29	27/263 = 0.10
EXER = 1				
5/278 = 0.02	13/276 = 0.05	66/325 = 0.20	56/179 = 0.31	140/1058 = 0.13
Marginal Proportions (TR360)				
7/342 = 0.02	14/367 = 0.04	77/388 = 0.20	69/224 = 0.31	

Table 1b: Proportion of DCS by P2 and EXER

P2 categories				
(2.99, 4.30]	**(4.30, 6.00]**	**(6.00, 7.80]**	**(7.80, 10.11]**	**Marginal Proportions (EXER)**
EXER = 0				
16/110 = 0.15	11/46 = 0.24	0/68 = 0.0	0/39 = 0.0	27/263 = 0.10
EXER = 1				
69/387 = 0.18	36/130 = 0.28	34/336 = 0.10	1/205 = 0.005	140/1058 = 0.13
Marginal Proportions (P2)				
85/497 = 0.17	47/176 = 0.27	34/404 = 0.08	1/244 = 0.004	

Next, we explore the influence of the explanatory variables on the *time to onset* of DCS, in hours. To do this, we estimate the survival and hazard rate nonparametrically (i.e., ignoring all covariates). For the survival curve, we use Breslow's nonparametric estimator modified for adjustment for ties by Fleming and Harrington (1984). For each recorded DCS time, t_i, define $d\bar{N}(t_i)$ to be the number of DCS cases in the data set with that recorded time, and

define $\bar{Y}(t_i)$ as the number of subjects in the risk set at time t_i. Also, define $d\hat{\Lambda}(t_i) = d\,\bar{N}(t_i)/\bar{Y}(t_i)$. Then, Breslow's estimator of the survival curve for the DCS data is

$$\hat{S}_B(t) = \prod_{j:\, t_j \le t} \exp[-d\hat{\Lambda}(t_j)] \tag{3.1}$$

(Therneau and Grambsch, 2000, p. 14-16). The adjustment for tied event times replaces $d\hat{\Lambda}(t_i)$ with $\sum_{j=0}^{d\,\bar{N}(t_i)-1}\left[\bar{Y}(t_i) - j\right]^{-1}$. This estimator (with adjustment for ties) is very similar to the familiar Kaplan-Meier estimator of survival. Our reason for using this estimator instead of the Kaplan-Meier estimate pertains to the method we use to deal with ties when fitting a Cox model. That is, the same method for dealing with ties is used in the Cox model fitting.

For the nonparametric hazard estimate, we use the Nelson-Aalen estimator (Collett, 1994, p. 28). This hazard estimate is constant in the interval $t_{(j)} \le t < t_{(j+1)}$ and for event times $t_{(j)}$ and $t_{(j+1)}$, and it estimates the risk of DCS per unit time in this interval. The hazard estimate is computed as

$$\hat{h}(t) = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})}, \qquad t_{(j)} \le t < t_{(j+1)}, \quad j = 1,\ldots,r-1 \tag{3.2}$$

where d_j is the number of DCS events at time $t_{(j)}$, and n_j is the number at risk for getting DCS at that time. The hazard is not estimated for the interval beginning at the last observed event time. Thus, the hazard estimate is truncated at the last observed event time in the data set.

Figures 1a and 1b plot several survival curve estimates. The first panel in Figure 1a shows the overall estimate of survival by hour as a dark solid line, along with 95% point-wise confidence intervals (as dotted lines) computed on the log survival using Greenwood's variance estimate (Therneau and Grambsch, 2000, p. 16). The 'X's denote censored times that do not coincide with DCS times in the data set (there is substantial overlap of censored times). The second panel stratifies the estimate by whether exercise was done at altitude. The first panel in Figure 1b stratifies the estimate by quartiles of TR360, and the second panel stratifies the estimate by quartiles of P2. For clarity, confidence intervals are excluded from the plots that have stratified estimates.

In the right panel of Figure 1a, the crossing of the survival curve estimates (see point 'A') indicates that the observed incidence of DCS was greater after one hour for those who exercised than for those who did not exercise at altitude. Prior to one hour, incidence of DCS was slightly higher for those who did not exercise. This may imply that the corresponding actual hazard functions for each exercise group are not proportional. However, point-wise confidence intervals on the two survival estimates (not shown) are wide enough to question whether an observed crossing is real. If the survival curves do indeed cross, this has implications for fitting traditional proportional hazards models such as the Cox (1972) model and certain parametric survival models that assume that the true hazards are proportional across covariate groups.

The right panel in Figure 1b shows that as TR360 increases, the survival probability at any time point decreases, as would be expected. P2 generally shows the opposite pattern. There is no strong evidence of crossing survival curves for either the TR360 categories or the P2 categories.

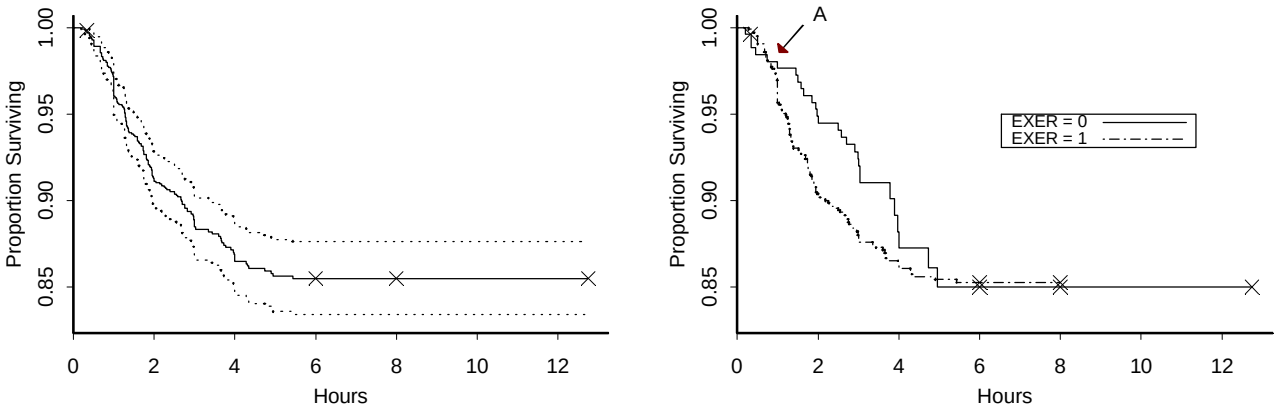

Figure 1a: Nonparametric estimates of survival for (left) all subjects (right) stratified by EXER. The point 'A' in the right panel denotes where the curves cross.

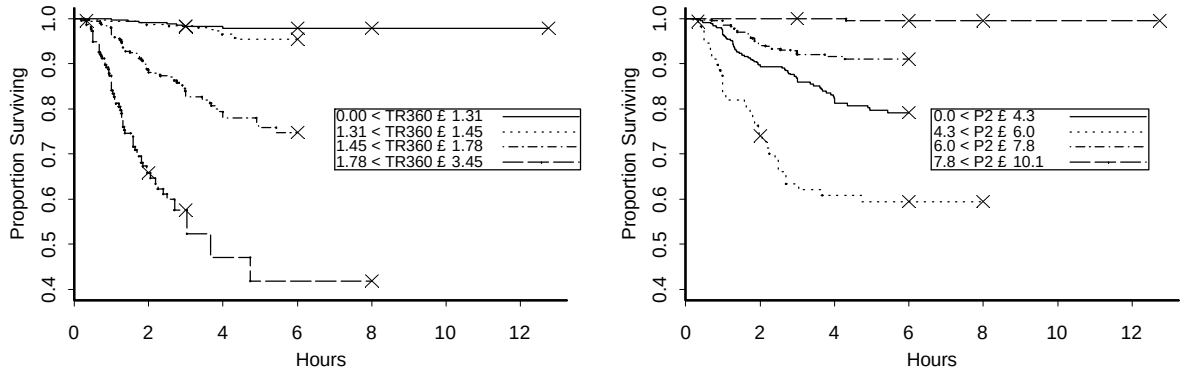

Figure 1b: Nonparametric estimates of survival stratified by (left) TR360 and (right) P2.

The survival curves in Figures 1a and 1b give information about survival probability over time, but present cumulative information. On the other hand, a hazard plot will show how the instantaneous risk of DCS changes by hour. For the same variables and categories used in Figure 1, we computed the estimated hazard function using (3.2). Because the graph of this estimate by time was very rough and difficult to interpret, we instead plotted the observed DCS times by the values obtained from (3.2), then fit a kernel density estimate over these values using the S-PLUS function *muhaz* (Hess et al., 1999). These smoothed curves were much easier to interpret. The curves are plotted in Figure 2.

Figure 2 shows that the estimated hazard for EXER = 1 increases much more rapidly than that for EXER = 0 and has its peak between one and two hours. The hazard rate declines after this and eventually is below the rate for EXER = 0 after about three hours. The hazard rate for EXER = 0 rises slightly, then is roughly constant between 2.5 and four hours before it sharply declines. There is no real indication of the crossing survival curves that were seen in Figure 1a at around the first hour. The slight crossing in the initial few minutes of the first hour disappears with the use of a higher local bandwidth in that area. Thus, this slight crossing may in fact be an artifact of random fluctuation.

The hazard estimates for the TR360 categories display the order that is expected based on their implied risk. The estimates for the first two categories starting from the smallest category are mostly flat. The hazard estimates for the two highest TR360 categories increase initially, then decline. (This pattern also appeared for the hazard estimate of the highest two PN2360 quartiles, which are not plotted). These patterns persisted even with higher bandwidths for the density estimates. Thus, risk of DCS is clearly related to TR360, but it may gradually have less impact over time.

The declining pattern appears across exercise conditions as well. But, the hazard rate estimate for a particular TR360 category differs depending on whether exercise is done at altitude (lower right panel). The hazard is much greater in the initial hours when exercise is done. Only the last two TR360 categories are shown in the figure for clarity, as the hazard functions are all very small for lower TR360. However, the kernel density estimate of the hazard rate is actually higher for EXER = 0 than for EXER = 1 when TR360 $\leq$ 1.31, at almost all time points.

Finally, the hazard estimates for P2 clearly show that the most hazardous category for P2 is between 4.3 and 6.0 psia.

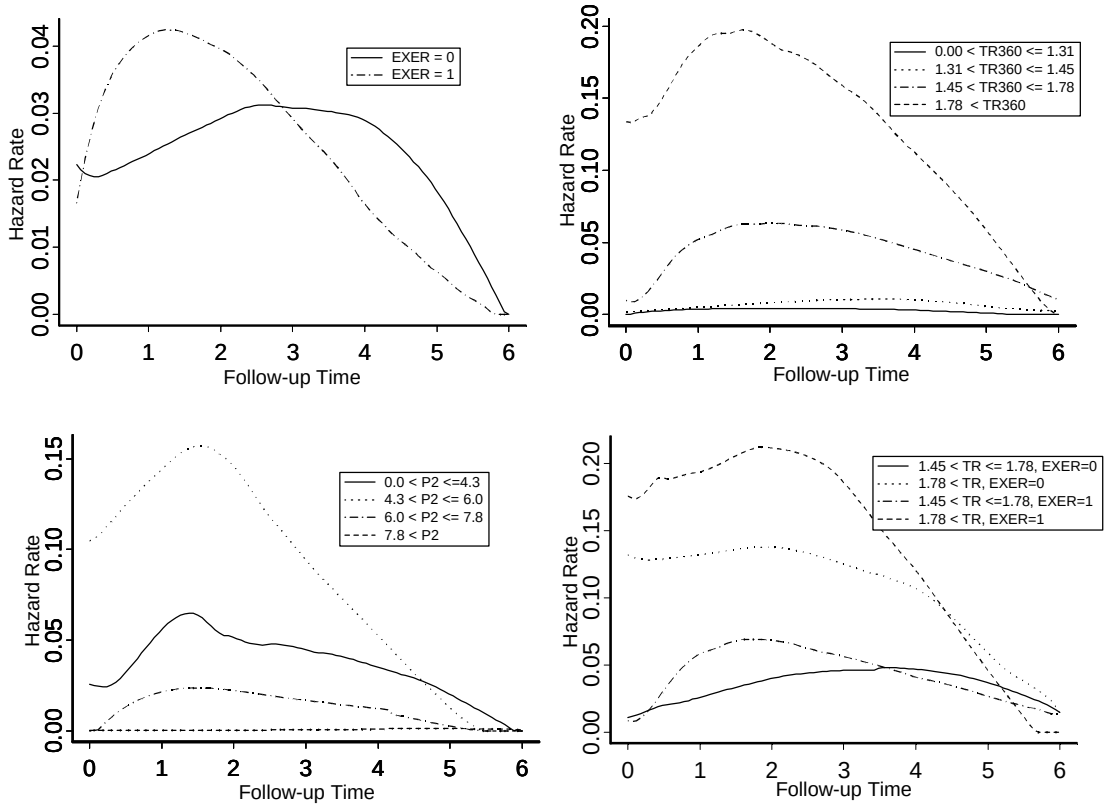

Figure 2: Estimated hazard plots by EXER, TR360 quartiles, and P2 quartiles.

Thus, Figure 2 shows some evidence against the PH assumption. It appears that the covariates may not act persistently on the hazard in that the estimated hazard rate for 'high risk' groups (e.g., EXER = 1 or high TR360) is not at a roughly constant distance from the hazard rates for other groups, and sometimes falls lower than that for lower risk groups after some time. Thus, some covariates may have less relevance on the hazard after a few hours. Also, the converging hazards of P2 and EXER by TR360 are not consistent with a PH assumption. However, the hazard estimates in Figure 2 are useful largely for exploratory purposes. We examine the assessment of PH in more detail in the next section, where we fit several Cox models and perform assessments based on the fit.

4. Assessment of Proportional Hazards

In this section, we assess whether hazards can be considered proportional (PH assumption) across all covariates. We use several graphical techniques to assess model assumptions and the fit of a potential Cox model that includes covariates from Table 1. For binary covariates, a comparison of nonparametric survival curve estimates may be suf-

ficient to decide PH because if the hazards were proportional, the survival curves for the two conditions would separate exponentially. The two curves would not cross each other. For example, Figure 1a showed a plot of estimates of survival probabilities stratified by exercise condition. The survival probabilities did not take into account the other covariates. The crossing curves at around hour = 1 indicated potential non-PH across exercise, barring any assessment of uncertainty. Non-PH would imply that the relative risk of DCS changes over time (hours exposed) for subjects who exercise versus subjects who do not exercise during exposure.

For continuous covariates it is not sufficient to rely only on stratified survival estimates to assess PH because the choice of stratification points is subjective and arbitrary in some cases. Thus, we need other alternatives to assess PH across the values of continuous covariates. One alternative is via the use of time-varying coefficients (Grambsch and Therneau, 1994). That is, one or more coefficients multiplying their respective covariates varies with time. If the coefficient multiplying a covariate is not constant over time, then the impact of that covariate on the hazard varies over time, leading to non-PH. If PH holds, a plot of the coefficient versus time will be a horizontal line. This plot is superior to the hazard estimate plots in Section 3 because we are not restricted to certain categories of continuous covariates. Also, we can perform formal tests for specific forms of departure from PH. The next subsection fits a Cox model. In it, we explain how the test of time-varying coefficients is conducted.

4.1. Initial Fit of a Cox PH Model

We fit the Cox PH model in (2.1) to the DCS data using partial maximum likelihood estimation. The S-PLUS function *coxph* was used to do the estimation, with ties handled via Efron's method (Therneau and Grambsch, 2000, p. 49, or S-PLUS, 2001, p. 390). This method of handling ties is similar to the adjustment used for the nonparametric survival estimate (Therneau and Grambch, 2000). We also included the covariates in Table 1, with the exception of PN2360. The fit of this model is shown in the Model 1 column of Table 2 with approximate standard errors in parentheses (obtained from the inverse of observed information, as computed from the partial likelihood). Thus, for two individuals differing by one unit in TR360 (all else equal), the individual with a higher TR360 has a higher expected risk by exp(2.142) = 8.52-fold. But, for two individuals differing by one unit in P2, the individual with lower P2 has a higher expected risk by 100*exp(0.307) % = 136%. Exercisers versus non-exercisers have approximately a 311% increase in risk.

Table 2: Maximum Partial Likelihood Estimates for Fitted Cox Models

	Model 1	**Model 2**
–2 Log LH	2120.01	2078.42
AIC = –2 Log LH + 2*p*	2126.01	2086.42
Parameter Estimates		
b_1 (TR360)	2.142 (0.201)	1.337 (0.271)
b_2 (PN2)	–0.307 (0.060)	–0.252 (0.067)
b_3 (EXER)	1.135 (0.265)	–3.663 (0.776)
b_4 (TR360:EXER)		2.481 (0.399)

In the Model 2 column of Table 2, we show the fit of a model with an interaction term in TR360 and EXER (TR360:EXER, in the table). Of the two models, Aikaike's Information Criterion (AIC) measures show that the interaction model (Model 2) fits better. With this model, the relative change in risk for repetitive exercisers versus non-exercisers depends on TR360. For TR360 greater than about 1.48, the predicted risk for exercisers is higher than for non-exercisers. Otherwise, the reverse is true. This relationship is consistent with the survival curves in Figure 1 and the hazard estimates in Figure 2 because risk is not uniformly higher for exercisers. In particular, as mentioned in the discussion about Figure 2, the kernel density estimate of the hazard rate for non-exercisers was

higher than that for exercisers when TR360 $\leq$ 1.31, for all time points. However, these results may only reflect the lower impact of exercise on DCS when TR360 is lower, instead of implying that no exercise is more hazardous in this TR360 range.

4.2. Test of Time-varying Coefficients in a Cox PH Model

To check whether the Cox model fit is valid, we must check the proportionality assumption. That is, we must check whether the effects of covariates on risk remain constant over time.

To illustrate the test of time-varying coefficients, we first describe the Schoenfeld (1982) residual. To do this, we use the notation of Therneau and Grambsch (2000). Let $t_1,...,t_d$ be the *d* unique ordered event times, and let $X_i(s)$ be the $p\times 1$ covariate vector for the *i*th individual at time *s*. For time-fixed covariates, this is just X_i. Also, define the 'weighted mean' of the $X_i(s)$ over those still at risk at time *s* as

$$\bar{x}(\hat{b},s)=\frac{\sum Y_i(s)\exp(X_i(s)\hat{b})\,X_i(s)}{\sum Y_i(s)\exp(X_i(s)\hat{b})} \tag{4.1}$$

where $Y_i(s)$ is the predictable variation process indicating whether observation *i* is at risk at time *s*, so that $Y_i(s)=1$ if observation *i* is still at risk at time *s* and is zero otherwise. The estimate $\hat{b}$ comes from fitting a Cox PH model. Then, a Schoenfeld residual is a $p\times 1$ vector that is defined at the *k*th event time as

$$s_k=\int_{t_{k-1}}^{t_k}\sum_i\left[X_i(s)-\bar{x}(\hat{b},s)\right]d\,N_i(s) \tag{4.2}$$

where $N_i(s)$ is a counting process that counts the number of events for observation *i* at time *s*. Thus, s_k sums the quantities $X_i(t_k)-\bar{x}(\hat{b},t_k)$ over observations that have experienced the event by time t_k. With no tied event times, the *k*th Schoenfeld residual is the sum of contributions to the derivative of the log partial likelihood by subjects who have experienced events by t_k (Hosmer and Lemeshow, 1998).

A *scaled* Schoenfeld residual is (4.2) divided by an estimate of its standard deviation. Therneau and Grambsch (2000) show that the standard deviation is the square root of the weighted variance of $X_i(s)$ at time *s*, where the weights are the same as in (4.1). The scaled Schoenfeld residuals are used in a test of proportional hazards.

For the *j*th covariate, Grambsch and Therneau (1994) express a time-varying coefficient as

$$b_j(t)=b_j+g_j g_j(t) \tag{4.3}$$

where $g_j(t)$ is a specific function of time. They show that the scaled Schoenfeld residuals have, for the *j*th covariate, a mean at time *t* of approximately $g_j g_j(t)$. Thus, a plot of the scaled Schoenfeld residuals by the event times may assess whether the coefficient g_j is zero or not, and what the function $g_j(t)$ might be. A linear regression line can also be fitted to the plot along with a test for zero slope. A nonzero slope is evidence against PH. As an alternative method of plotting, we can add the estimate of the regression coefficient to the scaled Schoenfeld residual to get a plot of the regression coefficient (as in (4.3)) by time. We have done this in Figure 3.

Figure 3 shows scatter plots of the scaled Schoenfeld residuals by time for each single covariate from Model 2 of Table 2. The smoothed curve in the plot is a natural spline with four degrees of freedom. The curve gives an indi-

cation of the path of the regression coefficient for that covariate by time. Ninety-five percent confidence bands are also given by dotted lines, using the variance of the estimated spline curve (Therneau and Grambsch, p. 134-135). The horizontal line is the estimate of the coefficient from the Cox model in Table 2. The p-values in the right-hand corners come from a test of significant linear change in the coefficient over time (Therneau and Grambsch, 2000, p. 131-134).

Figure 3 indicates a small changing effect of P2 on the hazard, but it is not significantly linear at the 0.05 level. Also, the confidence intervals contain the fixed estimate from Table 2. The change appears to be due to a small set of residuals at higher time points. Figure 3 shows the coefficient becoming more negative over time. Thus, risk increasingly lowers over time with higher ambient pressure at final altitude.

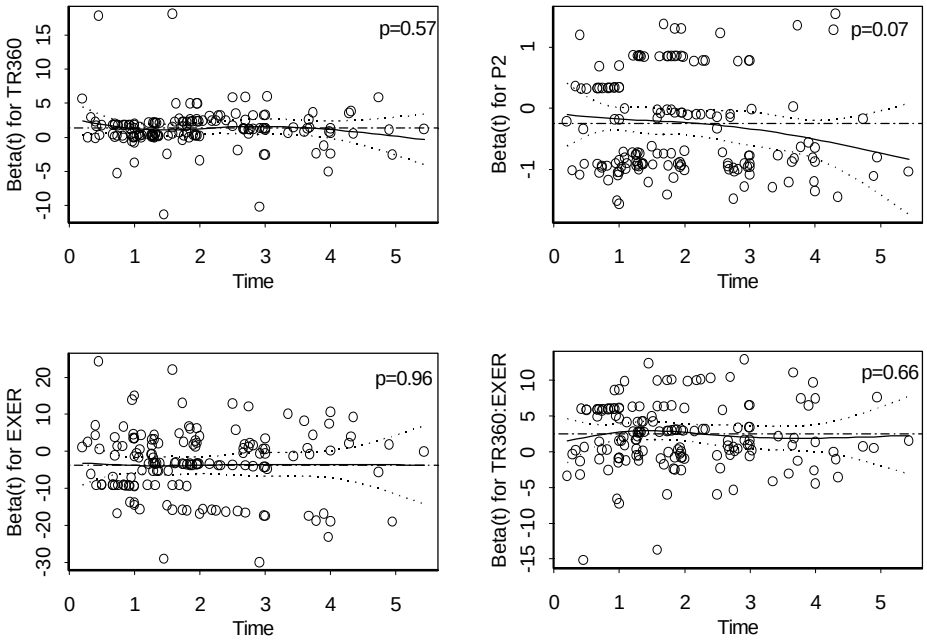

Figure 3: Smoothed scaled Schoenfeld residual plots for Model 2 (test of time-varying coefficients).

We can also use monotonic functions of time $g(t)$, such as log(*t*), on the abscissa. In this case, the p-value corresponds to a test of the addition of the time-dependent covariate $X * g(t)$ into the model, implying non-PH in *X* if the covariate is significantly different from zero (Therneau and Grambsch, 2000). Other specifications for $g(t)$ lead to various tests for PH in the literature. See Therneau and Grambsch (2000) for details.

We tried the transformation $g(t) = \log(t)$. Figure 4 shows scatter plots of the scaled Schoenfeld residuals by time on the log scale for each predictor term in Figure 3. None of the tests of log linear change in the coefficients are significant. However, many of the spline fits show clear nonlinear change, but these are due only to a few extreme points. For example, the time-varying coefficient for EXER appears to be strongly influenced by a few points at the low end of the time scale. In this case, exercisers have a higher hazard initially than non-exercisers. The hazard risk remains fairly constant throughout the rest of the study time. This pattern is consistent with a lessening of the importance of exercise on DCS over time.

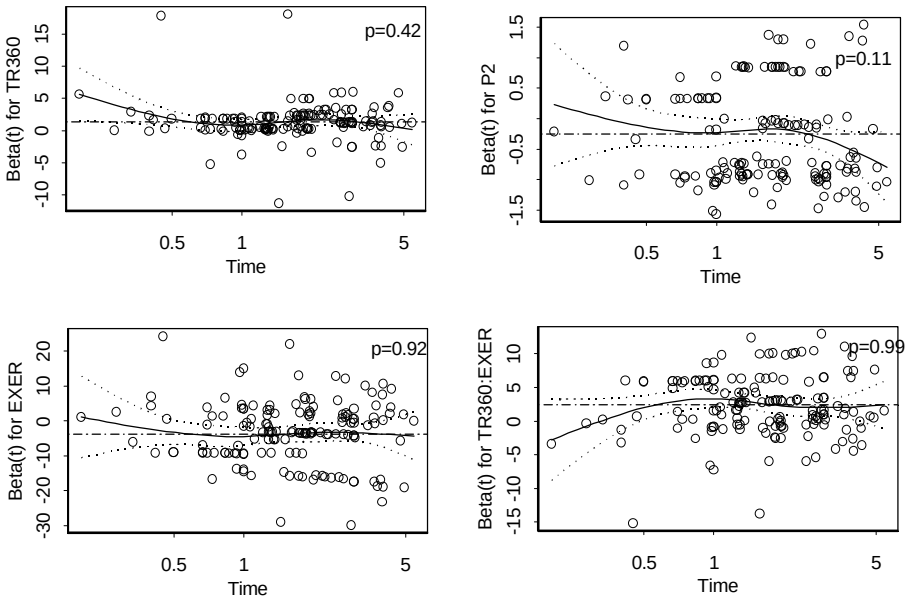

Figure 4: Test of time-varying coefficients for Model 2 using Log(Time) as the time transformation.

In a model excluding the interaction, we found significant evidence of time-varying coefficients for all of the variables. Hence, we conclude that Model 1 of Table 2 is not an appropriate PH model.

One problem with the significance test for a nonzero slope is that, because it is based on an ordinary least-squares line fit to the scaled Schoenfeld residuals, it is heavily influenced by outliers. Time transformations that are less influenced by outliers include rank time and the Kaplan-Meier (KM) transformation $g_j(t) = 1 - \hat{S}(t)$, where $\hat{S}(t)$ is the KM estimate (Therneau and Grambsch, 2000). (For DCS data, all p-values for the KM transformation were above 0.12.) Alternatively, we might use a rank correlation test or just rely on the smoothed spline fits to the scatter plots as a way to visualize non-PH, especially if a nonlinear trend were suspected. There is also a limitation in the form of non-PH detected by scaled Schoenfeld residuals. Complicated forms of non-PH that involve interactions between covariates and time-dependent coefficients (e.g., a different coefficient function for each value or set of values of a covariate) cannot readily be detected unless we suspect them and construct a Schoenfeld plot for that subset of values.

It is possible that evidence of time-varying coefficients appears because of other causes instead of non-PH. Therneau and Grambsch (2000) list some of these reasons, including omitted covariates and incorrect functional forms for covariates. We checked appropriate functional forms for the continuous covariates in the model (TR360 and P2) by substituting restricted cubic splines for each in the model. Plots of the fitted smoothing spline in P2 against the log hazard show that a quadratic term in P2 may be more appropriate than a linear term. A model with both linear and quadratic terms in P2 fits better (AIC = 2071.72) than a model with only a linear term. Thus, we include a quadratic term in the final model. When this term is included in the model, evidence of a time-varying coefficient is not as strong as that seen in Figures 3 and 4. Table 3 shows the coefficients of this model (called Model 3), and Figure 5 shows the results from tests of time-varying coefficients. We used mean deviations in P2 to reduce the correlation between estimated coefficients on the linear and squared terms. In Figure 5, D.P2 = $(P2 - \overline{P2})$.

Table 3: Maximum Partial Likelihood Estimates for Third Cox Model

	Model 3
–2 Log LH	2061.72
AIC	2071.72
Parameter Estimates	
b_1 (TR360)	1.893 (0.337)
b_2 $(\mathrm{P2}-\overline{\mathrm{P2}})$	–0.515 (0.122)
$b_3(\mathrm{P2}-\overline{\mathrm{P2}})^2$	–0.247 (0.072)
b_4 (EXER)	–2.096 (0.928)
b_5 (TR360:EXER)	1.583 (0.488)

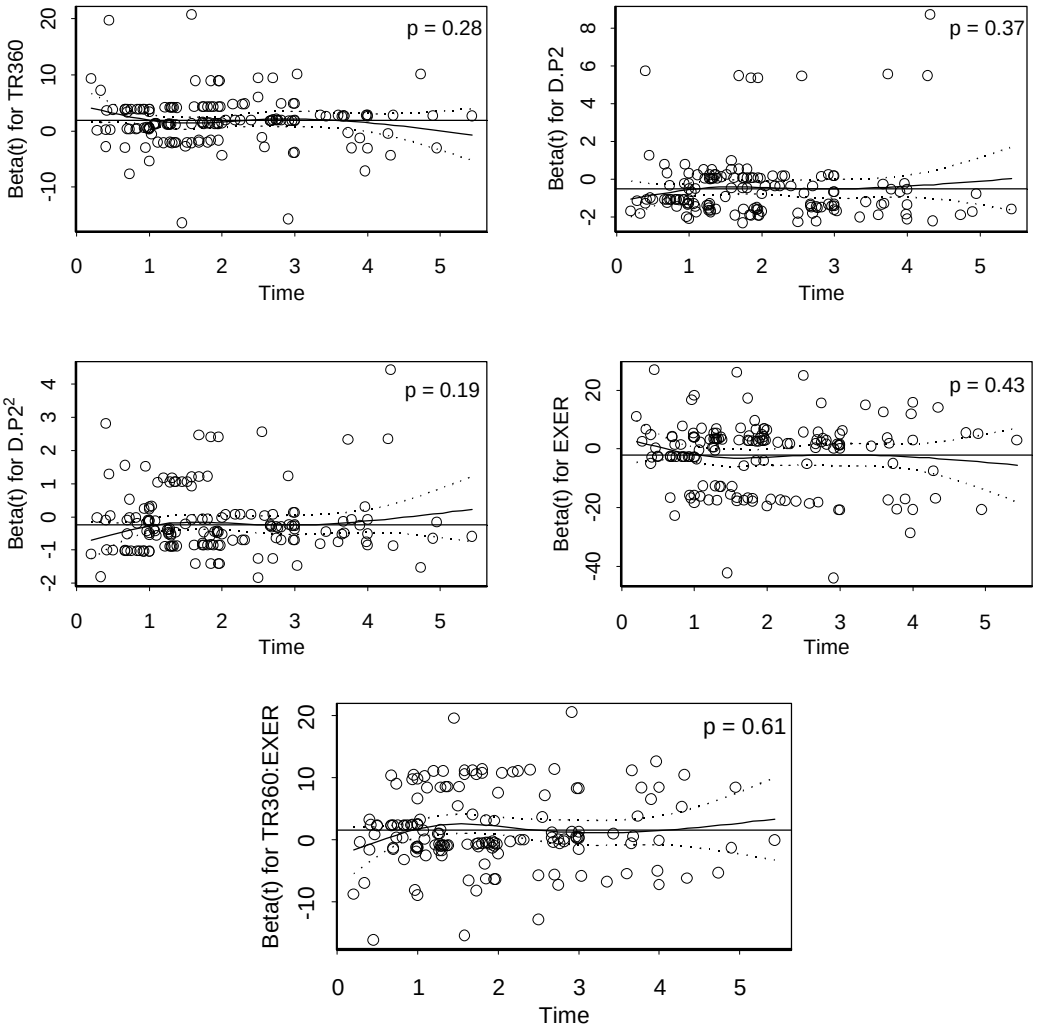

Figure 5: Test of time-varying coefficients for Model 3.

Omitted covariates are always a possibility and can cause non-PH (Therneau and Grambsch, 2000). The addition of a frailty term may account for unmodeled covariates. We briefly discuss this issue later when we mention frailty models in Section 7.

In the next subsection, we illustrate some graphical methods for testing PH after having fit a Cox model.

4.3. Graphical Tests for PH After Fitting a Cox Model

The two tests we apply in this section are the Andersen plot (Andersen, 1982, Klein and Moeschberger, 1997) and the Arjas plot (Arjas, 1988). The Andersen plot is based on the estimated baseline cumulative hazard function. The plot assesses each covariate for PH separately, given that the other covariates in the model satisfy PH. Continuous covariates must be categorized. First, a Cox model is fit using all covariates except the one being tested, and the fit is stratified on the covariate being tested. Denote the covariate being tested by the subscript *g*. Then, an estimate of the baseline cumulative hazard function (using the covariate data) is obtained for each stratum. Suppose there are *K* strata or categories of a covariate, and write $\hat{H}_{0_k}(t\,|\mathbf{x}_{(g)})$ as the estimated baseline cumulative hazard for the *k*th stratum, given all covariates except the *g*th (denoted by $\mathbf{x}_{(g)}$). The Andersen plot graphs $\hat{H}_{0_k}(t\,|\mathbf{x}_{(g)})$ versus $\hat{H}_{0_1}(t\,|\mathbf{x}_{(g)})$ for *k* = 2,…,*K* for all *t*. If the Cox PH model holds, the plotted curves should be straight lines through the origin, as the strata theoretically have proportional baseline hazards. The slope of the line should estimate the proportionality constant.

The estimate of the baseline cumulative hazard that we use is from Breslow (Klein and Moeschberger, 1997, p. 259). Let $t_{1_k},\ldots,t_{D_k}$ denote the distinct DCS times in the *k*th stratum, and d_{i_k} be the number of DCS cases from this stratum at time t_{i_k}. Then, Breslow's estimate is

$$\hat{H}_{0_k}(t\,|\mathbf{x}_{(g)}) = \sum_{t_{i_k}\le t} \frac{d_{i_k}}{W(t_{i_k};\hat{b})} \tag{4.4}$$

where $W(t_{i_k};\hat{b}) = \sum_{j\in R(t_{i_k})} \exp(\mathbf{x}_j^T \hat{b})$, $R(t_{i_k})$ is the risk set consisting of those subjects in stratum *k* still eligible to experience DCS at time t_{i_k}, and $\hat{b}$ is the MPLE from fitting a Cox model to all observations.

Figure 6 shows the Andersen plots for two covariates, P2 and EXER. On each graph is superimposed the 45-degree line for reference. P2 was categorized into quartiles. The first quartile (P2 $\le$ 4.3) was used as the first stratum, and is plotted on the abscissa. The times at which the estimated cumulative baseline hazards were computed ranged from zero to 13 hours, with increments of 0.05 hours. For the Andersen plot for EXER, the MPLE in the denominator of (4.4) did not include a strata-by-TR360 interaction, even though Model 3 included an interaction between EXER and TR360. The Andersen plot assumes the coefficients describing relative risk of covariates that satisfy PH do not change across strata.

In the left panel of Figure 6, the cumulative hazards for the second and third strata for P2 roughly follow the 45-degree reference line up to a value of around 0.13, indicating rough agreement among the cumulative hazards of the first three P2 strata for this range. The 'agreement' breaks down as the cumulative hazards increase further, however. The stratum with P2 > 7.8 had only one failure. Thus, its cumulative hazard is mostly at zero and is not shown. Regardless of which stratum is plotted on the abscissa, the same visual conclusions can be made.

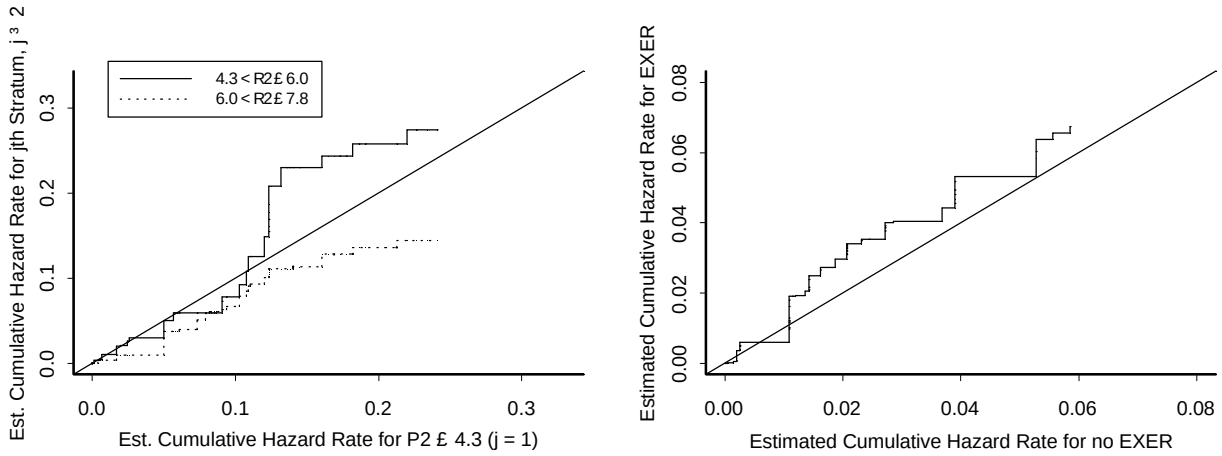

Figure 6: Andersen plots for assessing the PH assumption for Model 3 for (left) P2 and (right) EXER.

The Andersen plot for EXER in the right panel of Figure 6 shows a slightly concave pattern; thus, the hazard rate gap between EXER and no EXER may be slightly decreasing with time. This pattern was also seen in a plot of the ratio of the cumulative baseline hazards across EXER, estimated without covariates (not shown).

It is difficult to determine based on the Andersen plots in Figure 6 whether the PH assumption is violated. The graph in the right panel could be considered roughly linear through the origin. The same might be said for at least one of the curves in the left panel. Because these graphical tests for PH are illustrative and sometimes hard to interpret, Klein and Moeschberger (1997) suggest that several graphical assessments be compared. Thus, we also compute Arjas plots (Arjas, 1988) for each of the two covariates above. Below we give the basic idea of the plots. The appendix gives more details.

Arjas plots compare the observed cumulative number of failures to the estimated cumulative number of failures for each covariate being examined for PH. The estimated cumulative number of failures is derived from a model fitted to all covariates except the one being examined for PH. Each covariate to be examined is divided into *K* categories. An Arjas plot graphs the expected cumulative number of failures, $E_k(t_{i_k})$, by the observed cumulative number of failures $N_k(t_{i_k})$ at time t_{i_k} for the *i*th time in the *k*th category. Klein and Moeschberger (1997) give some guidelines for interpretation of the plot. If the covariate does not belong in the model, a plot of $N_k(t_{i_k})$ by $E_k(t_{i_k})$ should be close to a 45-degree line through the origin. If the covariate does belong in the model, the Arjas plot will give graphs for each category that are approximately linear, but with slopes differing from one. If the covariate in question has a non-PH effect on the hazard rate, the graphs will differ nonlinearly from the 45-degree line. Certain types of departures from linearity can give an idea of the relationship between hazards across categories or strata. For example, if the actual model is one where each category or stratum has a separate baseline hazard function (a stratified Cox model), where the ratio of the baseline hazard for category *g* to category g' is increasing with time, the respective cumulative baseline hazard functions will be concave for category *g* and convex for category g'. This pattern will be reflected in the Arjas plot.

Figure 7 shows Arjas plots for the two covariates P2 and EXER, with 45-degree lines superimposed. For the left panel, the curves for each level of P2 grow linearly away from the 45-degree lines. (The curve for P2 > 7.8 contained one point because there was only one failure for that stratum. Thus, its curve is not shown.). Also, the curve for EXER = 1 in the right panel is roughly linear with a slope less than 1.0. Based on these graphs, it appears that both P2 and EXER should be included in the model. In addition, there may be a non-PH effect of exercise on the hazard rate. There may be a need for different baseline hazards across EXER conditions because of the concave-like curve for EXER = 0. Thus, the ratio of EXER to no EXER baseline hazards may be decreasing with time.

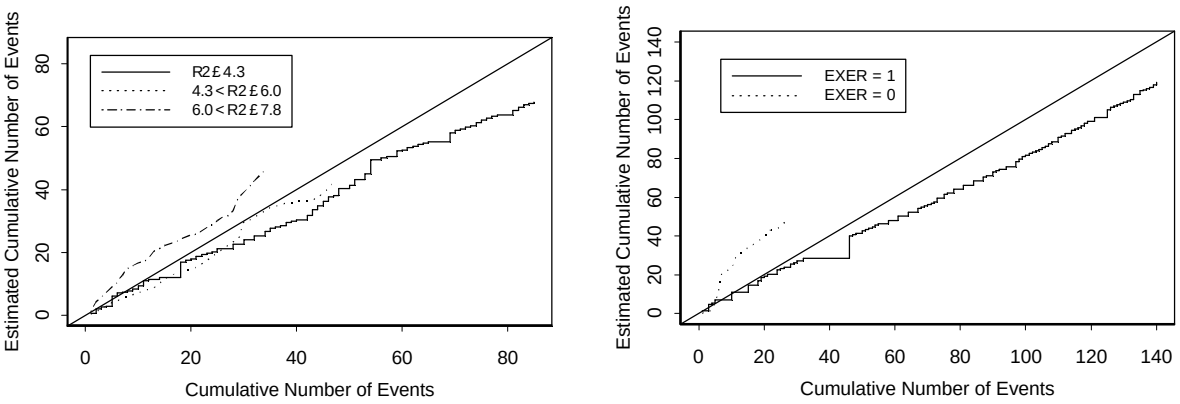

Figure 7: Arjas plots for assessing the PH assumption for Model 3 for (left) P2 and (right) EXER.

Based on the above assessments, we can conclude that the hazard rate may not be proportional over time across categories of some covariates. Specifically, the effects of covariates on the hazard rate may change over time. There are several options for attempting to correct non-PH or to be used as alternatives to a PH model. One alternative is to partition the time axis into sections where PH holds. Thus, to handle possible non-PH across EXER, we could partition the time axis at about one hour and only use failures after one hour in a Cox model. With so much censoring, however, this option is not viable for DCS data. Another option is to use an accelerated failure time (AFT) model. Therneau and Grambsch (2000) show these models can be detected by the time-varying coefficient tests mentioned in this section. AFT models are most appropriate in settings in which the time scale of the hazard function is either slower or faster (multiplicatively) than the time scale on which the measurements are made, as the covariates act by expanding or contracting time by a factor, $\exp(\mathbf{X}b)$. Another alternative to using the traditional Cox model is to stratify the model across levels of one or more covariates. We then assume that PH holds within each stratum. Two candidate covariates on which to stratify are EXER and P2, since these variables showed the most evidence of non-PH. In the next section, we describe the stratified Cox PH model. Then, we fit two stratified models and check PH within each stratum.

5. Stratified Cox Proportional Hazards Model

5.1. Fit of Stratified Cox Proportional Hazards Models

Two stratified versions of the Cox PH model were fit to the DCS data. In the first model, we stratified on EXER conditions. In the second model, we stratified on quartiles of P2. Stratification entails fitting separate baseline hazard functions across strata. A baseline hazard function represents the hazard rate over time for an individual with all modeled covariates set to zero. With a stratified Cox model, a proportional hazards structure does not necessarily hold for the combined data, but is assumed to hold within each stratum. However, the coefficients on the included covariates are common across strata so that the relative effect of each predictor is the same across strata, unless there is a significant strata-by-covariate interaction, which means that the effect of the covariate differs within strata. The estimated coefficients of a stratified Cox model are computed using the entire data set.

One disadvantage of using a stratified model is that an effect of the stratification covariate cannot be estimated in the model, at least in the usual sense of a coefficient estimate. This is a limitation if the stratification covariate is not merely a 'nuisance' variable that is recorded, but is of no substantive interest for the study (e.g., the clinic or hospital name at which recordings were made). However, if a model has been stratified on an important continuous variable that has been categorized, it is possible to also include the continuous variable in the model and thus estimate a relative effect for that covariate. The relative effect of the covariate is assumed to be the same within each stratum. This will be done for the continuous covariate P2. In addition, the baseline hazard function within each stratum can also be estimated using; for example, Breslow's estimate analogous to (4.4).

Table 4 shows the coefficient estimates of two stratified Cox models fit to the DCS data. Standard errors in parentheses were computed using the robust 'sandwich' variance estimator of Lin and Wei (1989) and implemented as an option in the S-PLUS function *coxph*. The first model (Model 4) is stratified on EXER and includes all covariates from Model 3 with the exception of EXER. It also includes a coefficient for the interaction between EXER strata and TR360, implying that the relative effect of TR360 on the hazard differs within each EXER stratum. This coefficient is indicated by b_5 (TR360:EXER=1) in the table. This notation implies that the term b_5 is only present in the prognostic index, $\exp(\mathbf{X}b)$, for exercisers. That is, $100\exp(b_1 + b_5)\% = 100\exp(0.718 + 2.71)\% \approx$ 3,081% represents the increase in risk per unit of TR360 for exercisers, and $100\exp(b_1)\% = 100\exp(0.718)\%$ $\approx 205\%$ represents the increase in risk per unit of TR360 for non-exercisers. Thus, there is a substantial increase in risk if one is exercising repetitively at altitude.

The second model (Model 5) stratifies on quartiles of P2, but it also includes P2 as a continuous covariate, as it appears in Model 3, as linear and quadratic terms. Model 5 purports that the quadratic effect of P2 is the same within the range of P2 values for each category. According to AIC, both models fit the data better than does Model 3. The model stratifying on P2 (Model 5) appears to fit better, although some of the standard errors are higher than they are for Model 4. Model 5 may be preferable because it allows a coefficient estimate for the relative effect of EXER on the hazard rate. Before examining this model more closely, we check the PH assumption again.

Table 4: Maximum Partial Likelihood Estimates for Stratified Cox Models

	Model 4 (stratified on EXER)	**Model 5 (stratified on P2)**
–2 Log LH	1916.90	1732.11
AIC	1924.90	1742.11
Parameter Estimates		
b_1 (TR360)	2.710 (0.241)	2.581 (0.504)
b_2 (P2$-\overline{\text{P2}}$)	–0.512 (0.129)	0.136 (0.336)
b_3(P2$-\overline{\text{P2}}$)^2	–0.246 (0.074)	–0.326 (0.110)
b_4 (EXER)		–2.450 (0.988)
b_5 (TR360:EXER)	NA	1.711 (0.532)
b_5 (TR360:EXER=1)	0.718 (0.247)	NA

5.2. Test of PH Assumption for Stratified Cox Models: Test for Time-varying Coefficients

The assumption of PH within strata can be checked using time-varying coefficients as in Section 4. However, the same limitations apply. That is, only certain types of non-PH can be detected. There is another complication in using the tests of Section 4. The variance used to compute the scaled Schoenfeld residuals is based on an overall estimate of the covariance of the covariates. The use of strata implies that the covariate patterns in the data may not be the same across strata and may definitely not be the same if there is evidence of strata-by-covariate interaction. Table 4 (Model 4) shows evidence for a significant strata-by-TR360 interaction, since the coefficient estimate is more than twice the magnitude of its approximate standard error. Thus, the computation of scaled Schoenfeld residuals requires a modification given in Therneau and Grambsch (2000, p. 142). The modification uses a within-stratum variance to compute the scaled Schoenfeld residuals within each stratum.

Figure 8 shows the result of the Therneau and Grambsch (2000) test for coefficients that vary linearly with time (i.e., time is represented on the identity scale). For the graphs that pertain to P2, the abbreviation D.P2 = (P2$-\overline{\text{P2}}$), as in Section 4.2. The p-values for the tests appear in the upper-right corners, and dotted-dashed horizontal lines

show the MPLEs from Table 4. Smoothed spline fits and their 95% confidence intervals are superimposed. The point-wise confidence intervals, which are computed automatically by the S-PLUS function *plot.cox.zph*, assume that the variance of the Schoenfeld residuals is constant over time. To check this assumption, we also computed the confidence intervals at each unique event time using the method in Grambsch and Therneau (1994). The two sets of confidence intervals were almost identical. We present the confidence intervals from the latter method in Figure 8.

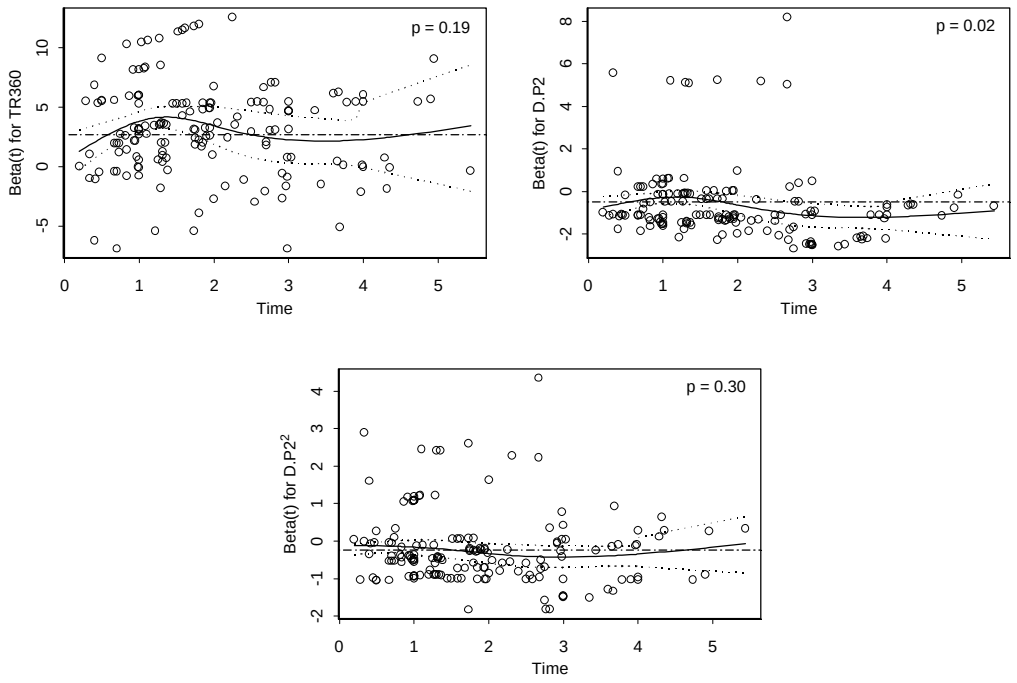

Figure 8: Test of time-varying coefficients when stratifying on EXER: smoothed scaled Schoenfeld residuals plot.

The p-value for D.P2 shows significant linear change, but the fitted spline through the residuals shows little variability. On the other hand, the p-value for TR360 is nonsignificant, but the smoothed spline fit shows a slight curvature. The magnitude of the slope was about 0.08 for both variables. The difference in significance is due to the relative variability of residuals for each covariate. The fact that the MPLE is not always contained within the given 95% confidence intervals on the spline fit may be indicative of a changing coefficient with time. Thus, stratifying only on EXER does not seem to have removed non-PH. The relative effect of TR360 appears to increase initially over time before it decreases after about 1.5 hours.

Figure 9 shows the result of the Therneau and Grambsch (2000) test after stratifying on quartiles of P2. In Model 5, we also include P2 as a continuous covariate. But, Figure 9 does not reflect a model with P2 as both a stratifying variable and a covariate in the model. This will not likely affect the test because the term $g_j g_j(t) \times \mathrm{P2}$, using the notation in equation (4.3), could just be absorbed into the baseline hazard function for the particular stratum of P2.

None of the p-values in Figure 9 shows significant linear change, and the fitted splines through the residuals show little curvature. Furthermore, the MPLEs are contained within the 95% confidence intervals on the spline fit. Plots using log time, as well as other simple monotonic functions of time, showed relatively similar results.

In the next subsection, we further assess PH for a model that stratifies on P2.

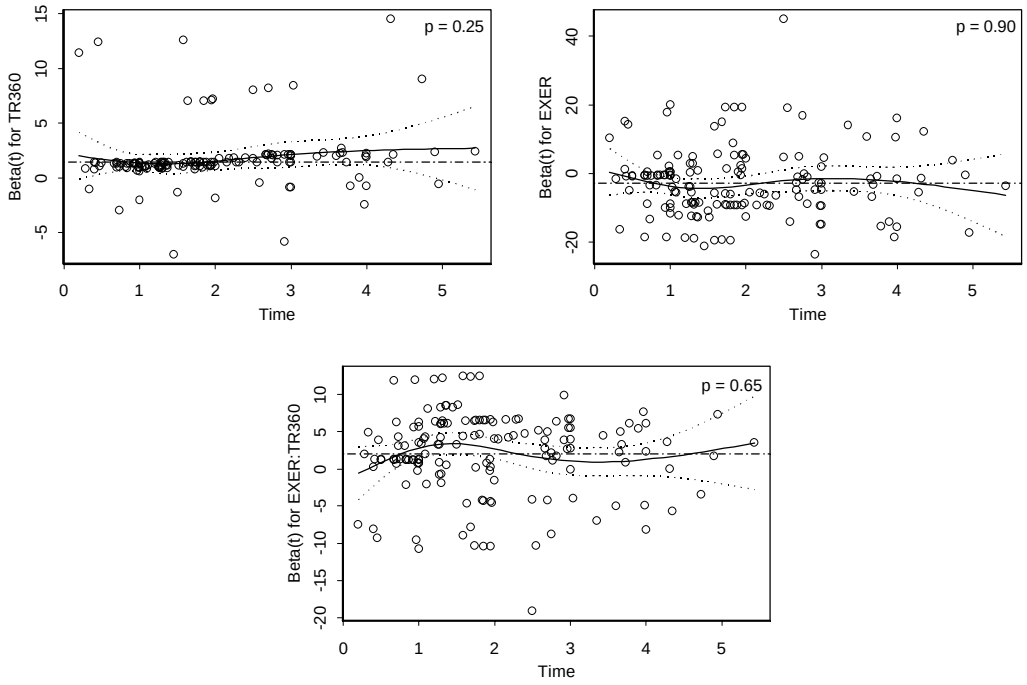

Figure 9: Test of time-varying coefficients when stratifying on P2 (Model 5): smoothed scaled Schoenfeld residuals plot.

5.3. Graphical Tests for PH After Fitting a Stratified Cox Model

Andersen plots and Arjas plots are easily modified for construction with stratified models. For a model stratifying on quartiles of P2 (Model 5), an Arjas plot can be constructed for assessing whether EXER should be stratified as well. The resulting plot is shown in Figure 10. The plot looks better than the analogous plot in Figure 7 for the EXER = 0 condition, as some of the nonlinearity is removed. However, the curve for EXER = 1 is slightly more nonlinear.

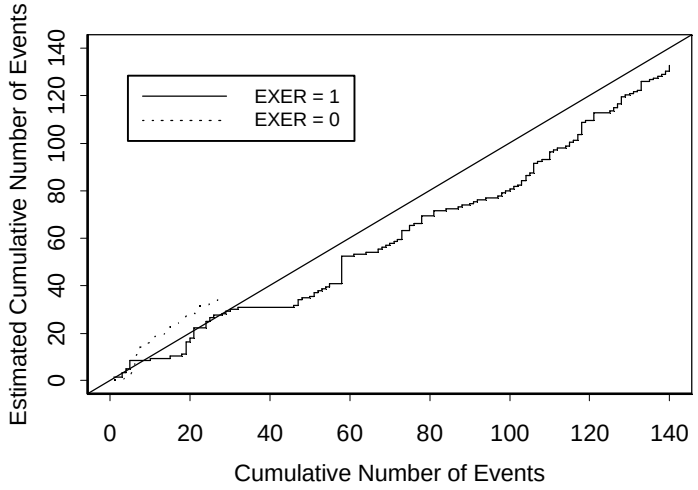

Figure 10: Arjas plot for assessing the PH assumption for Model 5 for EXER.

The corresponding Andersen plot in Figure 11 for EXER looks much more linear than the analogous plot in Figure 6. Thus, the cumulative baseline hazards appear to be roughly equal for the EXER = 1 and EXER = 0 conditions.

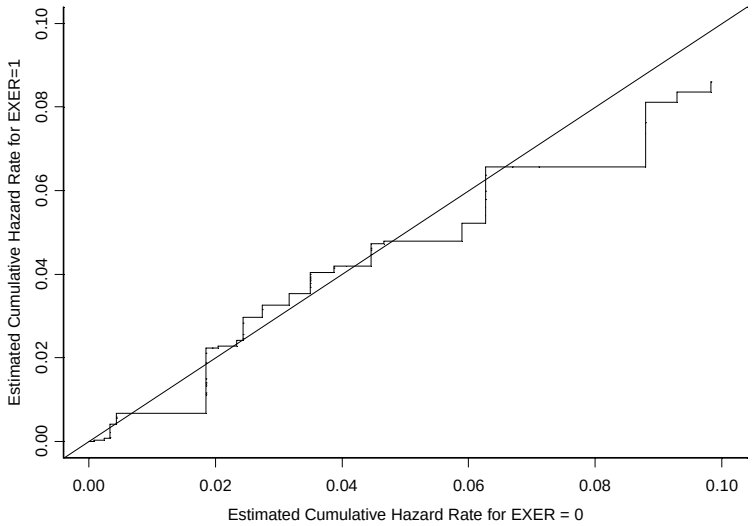

Figure 11: Andersen plot for assessing the PH assumption for Model 5 for EXER.

5.4. Expected Survival from a Stratified Cox Model

The expected, or predicted, survival curve based on Model 5 was computed for hypothetical individuals with certain values on the explanatory variables. The basic estimated expression is the survival curve in (2.2), with the MPLEs substituted for b and an estimate of the baseline survival substituted for $S_0(t) = \exp(-H_0(t))$, where $H_0(t)$ is the baseline cumulative hazard function. The method of estimation for $H_0(t)$ was the Fleming-Harrington (F-H) method (Therneau and Grambsch, 2000, p. 267), implemented as an option in the S-PLUS function *survfit*. This method is similar to using Breslow's estimate for $H_0(t)$, but it is appropriate for tied data. There was virtually no difference between the F-H and Kalbfleisch-Prentice methods. The reason for choosing the F-H method is that it deals with tied events in the same way that Efron's approximation deals with tied events in obtaining the partial maximum likelihood estimates (Therneau and Grambsch, 2000).

The left panel of Figure 12 shows expected survival probabilities for the midpoints of the first three quartile categories of P2, with remaining covariates TR360 = 1.60 and EXER = 1. (The fourth P2 category was omitted because there was only one failure, and the estimate of its baseline cumulative hazard function was therefore unreliable). The right panel of Figure 13 uses a TR360 value of 1.75. Confidence intervals are omitted for clarity. For both TR360 values, the predicted survival probabilities are roughly the same for the first hour and a half. After that, the second category ($4.3 \leq$ P2 < 6.0) is the most risky regardless of TR360 value.

To the extent that the expected survival can be compared with the nonparametric estimates in Figure 1b, predicted survival appears somewhat more consistent with the nonparametric estimates when the TR360 value is higher. Interestingly, the first P2 category is predicted to survive longer than the third category, whereas the nonparametric estimate claims the reverse. However, the nonparametric estimates are not constructed at specific covariate values. Thus, a direct comparison between Figure 12 and Figure 1b is unwise. If a value higher than the midpoint of each P2 category is used for the hypothetical individual, the third P2 category is expected to survive longer than the first category. Note, however, that the highest P2 category is not necessarily automatically the least risky of the four categories because it is not low ambient pressure per se that contributes to DCS. It is how ambient pressure compares to nitrogen partial pressure that is important in the contribution to DCS.

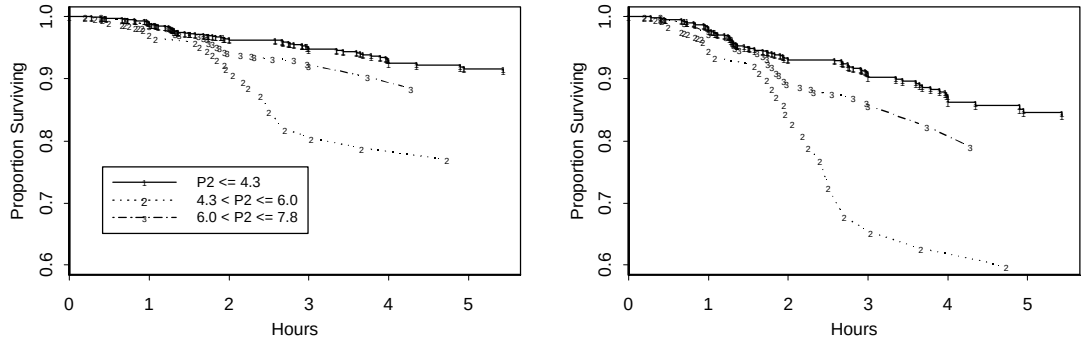

Figure 12: Expected survival for hypothetical individuals who exercised at altitude with left: TR360 = 1.60 and right: TR360 = 1.75.

Figure 13 shows expected survival for non-exercisers, with TR360 values of 1.60 and 1.75 for the left and right panels. The difference between Figure 13 and Figure 12 is a matter of degree, as would be expected from the model that was fit.

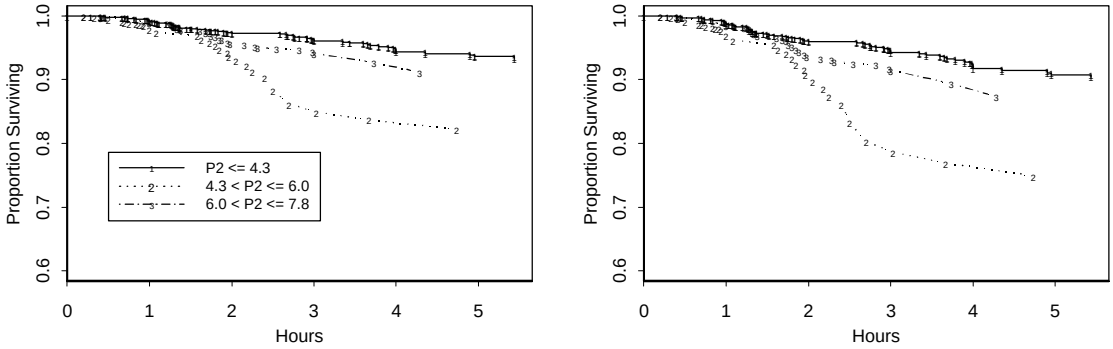

Figure 13: Expected survival for hypothetical individuals who did not exercise at altitude with left: TR360 = 1.60 and right: TR360 = 1.75.

In the next section, we discuss aspects of lack of fit in a model stratified on P2 quartiles.

5.5. Lack of Fit in the Stratified Cox Model

Residuals for the Cox model are not as useful in directly assessing global model fit as are residuals from a linear model or another type of parametric model (Therneau and Grambsch, 2000, Chapter 4). However, certain types of residuals can be used for specific purposes. And, some authors use residuals upon which to base global assessments of goodness-of-fit of a Cox model (e.g., Parzen and Lipsitz, 1999). In this subsection, we discuss the use of residuals and global goodness-of-fit tests to assess the fit of the stratified Cox model.

5.5.1. Deviance Residuals and Normal Deviate Residuals for Assessing Poorly Predicted Individuals

The deviance residual can be used for identifying 'poorly predicted' individuals (S-PLUS, 2001). The deviance residual is a standardized version of the martingale residual. For fixed-time covariates, a martingale residual is defined for each individual by

$$\hat{M}_i = N_i - \exp(\mathbf{x}_i^T \hat{b})\, \hat{H}_0(\hat{b},\, t_i) \tag{5.1}$$

where N_i is the number of events for individual *i* at time t_i (the number will be either 0 or 1 for these data). The estimate of the cumulative baseline hazard function at time t_i, $\hat{H}_0(\hat{b},\, t_i)$ is Breslow's estimate in (4.4). The martingale residual has the interpretation of the observed number of events for each individual minus the conditionally expected number given the fitted model and the follow-up time [0, t_i]. Thus, martingale residuals measure 'excess events' in an individual. For example, an individual who 'died' earlier than expected by the model will have a positive residual. An individual who lived too long will have a negative martingale residual. As many authors note (e.g., Therneau and Grambsch, 2000; Klein and Moeschberger, 1997), the martingale residual is highly positively skewed. Thus, the deviance residual is used in its place. The deviance residual is considered a normalizing transformation because a one-term Taylor series expansion shows that it divides the martingale residual by the square root of 'expected number of events' (Therneau and Grambsch, 2000, p. 83). Deviance residuals scattered about zero in a plot indicate a good fit of the model (S-PLUS 2001, p. 355). As with similarly defined deviance residuals for generalized linear models, we might consider residuals exceeding about three in magnitude to belong to poorly predicted individuals. However, this criterion is arbitrary, as there is no reference distribution for deviance residuals (Therneau and Grambsch, 2000).

The left panel of Figure 14 shows a plot of deviance residuals by observation for Model 5. The observations are ordered by recorded DCS time. Observed DCS cases are marked with an 'o', and censored cases are marked with an 'x'. A solid horizontal line is at zero. All of the residuals with magnitude greater than three are observed DCS cases, but there are not many. The right panel of Figure 14 shows the same residuals plotted by value of the linear predictor, $\mathbf{x}_i'\hat{b}$ for Model 5. (The sloping pattern seen in the residuals is a consequence of the way deviance residuals are computed). Again, censored and uncensored cases are marked with 'x's and 'o's, respectively. A LOWESS curve with 25% span is fit to all the residuals. If the residuals were truly scattered about zero, the smooth curve would show no trend. There is a slight trend downward. The pattern of the residuals indicates that some low-to-moderate-risk individuals (those who have a linear predictor between about –2 and 0) may be getting DCS too early as compared to that predicted by the model.

Several authors have commented on the limited usefulness of deviance residuals (Therneau and Grambsch, 2000; Nardi and Schemper, 1999). Deviance residuals have no reference sampling distribution, and a normal approximation has been shown to not be a satisfactory approximation (Nardi and Schemper, 1999). Nardi and Schemper derive new residuals called *normal deviate residuals* that have a standard normal distribution conditional on the survival function being known. Normal deviate residuals are designed to detect outlying cases, which means individuals who 'died too early' or 'lived far too long' as compared to that predicted by the fitted Cox model.

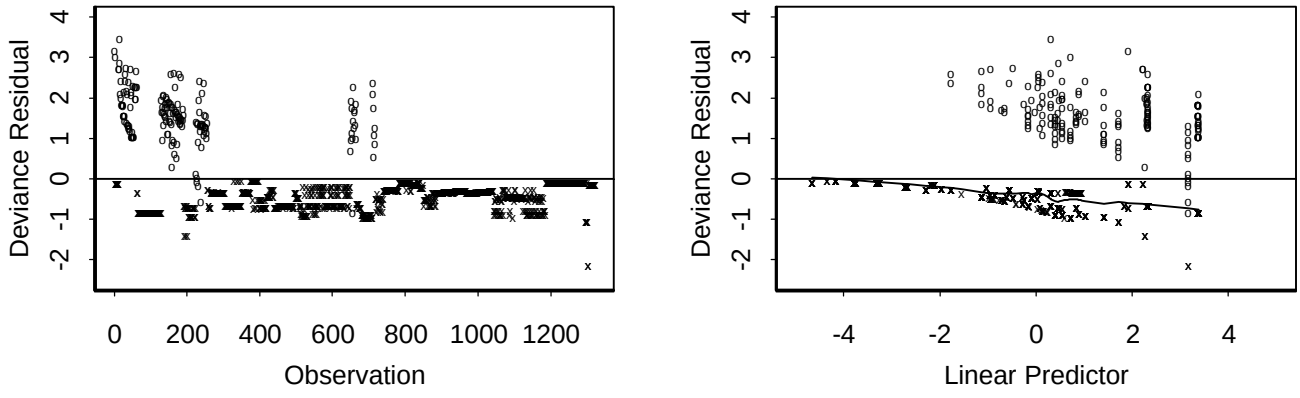

Figure 14: Deviance residuals for Model 5 plotted against (left) observation and (right) linear predictor. The x's represent censored observations and the o's represent observed DCS cases.

Nardi and Schemper (1999) regard prediction of survival by the Cox model to be perfect for an individual if that individual's observed survival time and estimated median survival time agree. To compare the observed survival time with the estimated median survival time, Nardi and Schemper compare the estimated survival probability at the event time with 0.5. Treating the survival probability for individual *i*, at time T_i, $S_i(T_i)$, as a success probability for a binomial model, they consider a probit transformation of $S_i(T_i)$, $N_i = \mathrm{F}^{-1}\{S_i(T_i)\}$, for a residual, called a *normal deviate residual* (because $\mathrm{F}(x)$ is the cumulative distribution function (CDF) of the standard normal distribution). Departures from perfect prediction result in a larger magnitude of the normal deviate residual. With no censored observations, assuming $S_i(\cdot)$ known, the sampling distribution for the normal deviate residual is the standard normal. When $S_i(\cdot)$ is replaced with an estimator (Nardi and Schemper use the Nelson-Aalen estimator for the baseline cumulative hazard), the resulting residual converges in probability to its estimand. For an uncensored observation, *i*, with the observed event at time t_i, the normal deviate residual is

$$n_i = \mathrm{F}^{-1}\{\hat{S}_i(t_i)\} \tag{5.2}$$

Residuals for censored observations are computed using the fact that the true survival time T_i is greater than the observed censored one t_i^C and, therefore, that the distribution of $S_i(T_i)$ will be uniform within $[0,\ \hat{S}_i(t_i^C)]$. The normal deviate residual for a censored observation is then the estimated conditional expected value of (5.2) given that $N_i \le \mathrm{F}^{-1}\{\hat{S}_i(t_i^C)\}$. Thus, the normal deviate residual for a censored observation is

$$n_i^e = -\frac{\exp(-0.5(n_i^C)^2)}{\sqrt{2p}\ \hat{S}_i(t_i^C)} \tag{5.3}$$

where $n_i^C = \mathrm{F}^{-1}\{\hat{S}_i(t_i^C)\}$ (Nardi and Schemper, 1999, Appendix A).

Nardi and Schemper (1999) use simulations to show the empirical distribution of the normal deviate residual approximates a standard normal much better than does a deviance residual. Thus, we could use standard normal percentage points as cutoffs for declaring an observation as outlying. However, because of the averaging process used to get the censored residuals, the empirical distribution is more concentrated than the theoretical normal distribution. Thus, for each censored residual, Nardi and Schemper compute the probability that, if uncensored, the residual would surpass the negative cutoff point (i.e., the observation would have lived too long). They show that this probability is $P_i = \min[1.0,\ a/\hat{S}_i(t_i^C)]$, where a is the 'cutoff' percent of the largest residuals in the negative direction. Thus, a censored individual with a cumulative survival probability less than a is definitely outlying. They recommend a threshold of 0.3 for P_i.

We computed normal deviate residuals for Model 5 and plotted them in Figure 15. The right panel shows the normal deviate residuals against the linear predictor. As with Figure 14, observed DCS cases are marked with an 'o' and censored cases are marked with an 'x'. Dotted horizontal lines denote the ±1.96 cutoff points from the theoretical normal distribution assuming a = 0.05. A solid horizontal line marks the zero point. All but two of the residuals exceeding the cutoff belong to uncensored observations. These observations experienced DCS too early, as compared to their estimated risk score based on the linear predictor. Two censored cases are outlying in the negative direction (both at –2.37). The P_i values for these two cases were 1.0. No other censored cases had P_i values greater than 0.3. The records for the two cases are shown below.

	DCS	censor	P2	EXER	TR360	PN2360
1302	8	0	4.4	1	2.28179	10.0399
1303	8	0	4.4	1	2.28179	10.0399

According to their covariate values, the high TR360 and EXER value of one puts both observations at high risk, yet they were censored at eight hours. Thus, they 'survived' already for too long.

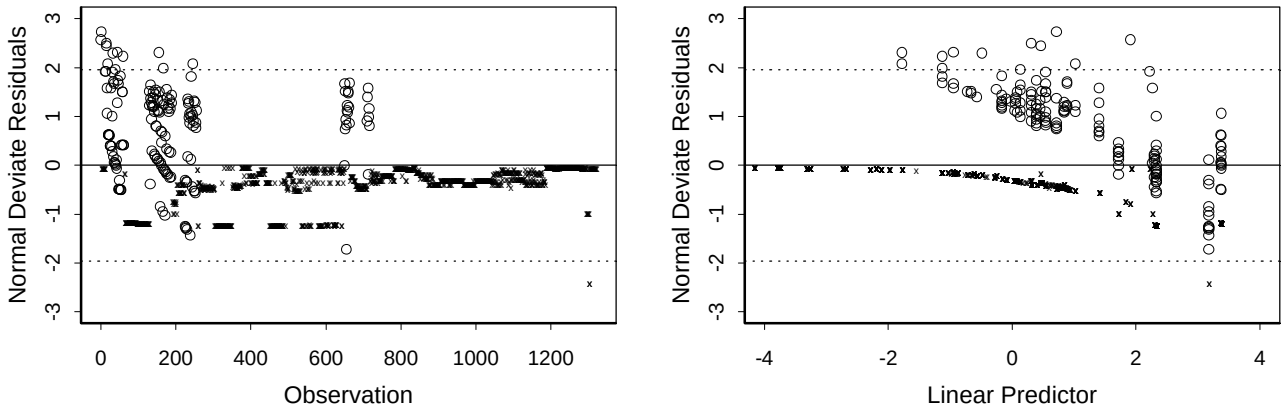

Figure 15: Normal deviate residuals for Model 5 plotted against (left) observation and (right) linear predictor. The x's represent censored observations and the o's represent observed DCS cases.

According to Figures 14 and 15, there are some observations that are poorly fit by Model 5. They tend to be uncensored cases that experienced DCS sooner than expected.

5.5.2. Global Goodness-of-Fit Using Martingale Residuals

A formal test of overall goodness-of-fit of the Cox model was proposed by Parzen and Lipsitz (1999) and independently by May and Hosmer (1998). The test compares observed and (model-based) expected numbers of events within covariate risk groups and computes a chi-square test. The test is similar to the Schoenfeld (1980) test, but suggests a partition of the covariate risk space that is more automatic. The covariate regions are defined by predicted risk scores, $\hat{y}_i = \exp(\mathbf{x}_i'\hat{b})$, where $\hat{b}$ is the MPLE from the fit of a Cox model. The cut-points of, say, *G* regions are defined by percentiles of the $\hat{y}_i$ values, called percentiles of risk, such that each category ideally contains roughly the same number of observations. Each observation is classified into one of these *G* categories depending on its risk score, and $(G-1)$ dummy variables are introduced into the Cox model. The score test of the resulting set of $(G-1)$ coefficients constitutes a significance test for overall fit of the Cox model. Sample size guidelines given by Parzen and Lipsitz follow those for general chi-square tests: In order for the score test to reliably have an approximate chi-square distribution, at least 80% of the categories must have estimated expected count of at least five and all estimated expected counts should exceed one. Expected counts were estimated using the estimated martingale residuals (at infinity) from the Cox model fit. The sum of the observed number of events minus the sum of the estimated martingale residuals within each category give the estimated expected count for that category (Parzen and Lipsitz, 1999; or May and Hosmer, 1998).

We performed this test for Model 5. We partitioned the risk scores into seven categories, as defined by break points given in Table 5, giving the indicated expected counts. The categories were chosen to achieve the expected sample size rules given by Parzen and Lipsitz (1999). To do this, we first calculated the septiles of the risk scores. Then, to get an expected count exceeding 1.0 in the first category, we raised the first category's upper cutpoint. Thus, not all categories have the same number of observations. The value of the score test was 7.80 with p = 0.253 (df = 6). Thus, we do not reject the hypothesis of model fit, at a significance level of 0.05. Furthermore, all other groupings we tried also did not reject the hypothesis of model fit at a significance level of 0.05. A comparison of the observed and expected events by risk group appears in Table 5. The two risk groups [1.370, 2.320) and [2.320, 10.284), representing roughly the 5th and 6th septile groups, are the worst predicted. Thus, although the model fits significantly well, there is room for improvement, particularly in the high risk area.

Table 5: Observed Failures and Expected Failures for Model 5

	Number of Obns in Region*	Observed Failures	Expected Failures
Risk group			
[0.00, 0.170)	263	2	1.23
[0.170, 0.518)	120	10	8.90
[0.518, 1.042)	193	13	13.09
[1.042, 1.370)	183	18	16.02
[1.370, 2.320)	187	29	38.48
[2.320, 10.284)	273	60	52.50
[10.284, 30)	102	35	36.78

*****Includes censored times.

Parzen and Lipsitz (1999) mention in passing that their test can be used as a formal test for PH by using time intervals as well as risk groups to divide the observations. However, in our experience this version of the test was very difficult to use correctly if we obeyed the sample size recommendations because it involved arbitrary decisions about the partitioning of the time-by-covariate space.

5.5.3. Assessment of Influential Observations

In general, for an observation to be influential on the fit of a model, it must be far from the mean of the covariates (high leverage) and have a large residual (Wilcox, 2001). In a Cox model, the residual is the martingale residual in (5.1), and the mean of the covariates changes over time as individuals leave the risk set (Hosmer and Lemeshow, 1998). Thus, leverage must be determined from a weighted average of the distances of the covariate values to the risk set means, where the risk set means are defined in (4.1). The so-called *score residuals* are these weighted averages that are defined for each observation for each covariate in the model (for further details, see Hosmer and Lemeshow, 1998; or Collett, 1994). To obtain a measure of influence of an observation on an MPLE of a coefficient, we scaled the score residuals by an estimate of the variance of the coefficient estimates. The resulting scaled residual, which is called the *scaled score residual* (or *dfbeta* residual), is used as a measure of influence (Therneau and Grambsch, 2000). The scaled score residual for the *k*th covariate and *i*th observation approximates the change in the *k*th coefficient estimate if the *i*th observation were removed from the data set and the model reestimated without that observation (see Therneau and Grambsch, 2000, for more details).

Figure 16 shows dfbeta residuals for Model 5. For each covariate, we have plotted the observation (in order of recorded DCS time) by the approximate scaled change in the coefficient after removing the observation from the model. This is the dfbeta residual divided by an estimate of the standard error of the coefficient. If the removal of an observation causes the coefficient to increase, the dfbeta residual is negative and vice versa. Figure 16 shows that although the influence values are much higher for some observations than others (mostly corresponding to uncensored cases), none of the observations exerts a change greater than about 30% of a standard error. For the covariate with the largest standard error (EXER), no observation exerts a change greater than about 20% of a standard error (about 0.20 in magnitude). Interestingly, the same two observations that had 'large' normal deviate residuals also had the highest influence magnitudes (–0.325) on the coefficient for TR360 × EXER. These individuals may be considered unexpected 'long-term survivors.' Long-term survivors tend to have a large effect on the MPLEs of regression coefficients in a Cox model (Valsecchi et al., 1996).

If the presence of influential points causes concern, we may obtain robust estimators of the coefficients using a weighted partial likelihood. Schemper (1992) and Valsecchi et al. (1996) present weighted partial likelihood estimates where weights are used on the contribution of each event time to the log likelihood. Valsecchi et al. used a weighted partial maximum likelihood estimation to get estimates of the coefficients that limit the influence of long-term survivors, whereas Schemper used it in the presence of non-PH for one or more covariates. Both authors use similarly defined weights based on the Kaplan-Meier survival estimate, but Valsecchi's development is specific to stratified Cox models. We followed Valsecchi's method for DCS data, and used weights at each event time equal to

the Kaplan-Meier estimate of survival within each P2 stratum (giving more weight to early failures in the stratum). The resulting coefficient estimates and approximate standard errors in Table 6 (Model 6) show that the coefficient estimates are similar except for the coefficient describing the relative effect of D.P2, which is now negative rather than positive. Thus, the quadratic effect of P2 is of the same shape but is not as pronounced for P2 values less than the mean (6.20), as it was in Model 5. As P2 passes its mean of 6.20 psia, the linear effect on the log hazard is positive for Model 6, whereas for Model 5 this linear effect was negative.

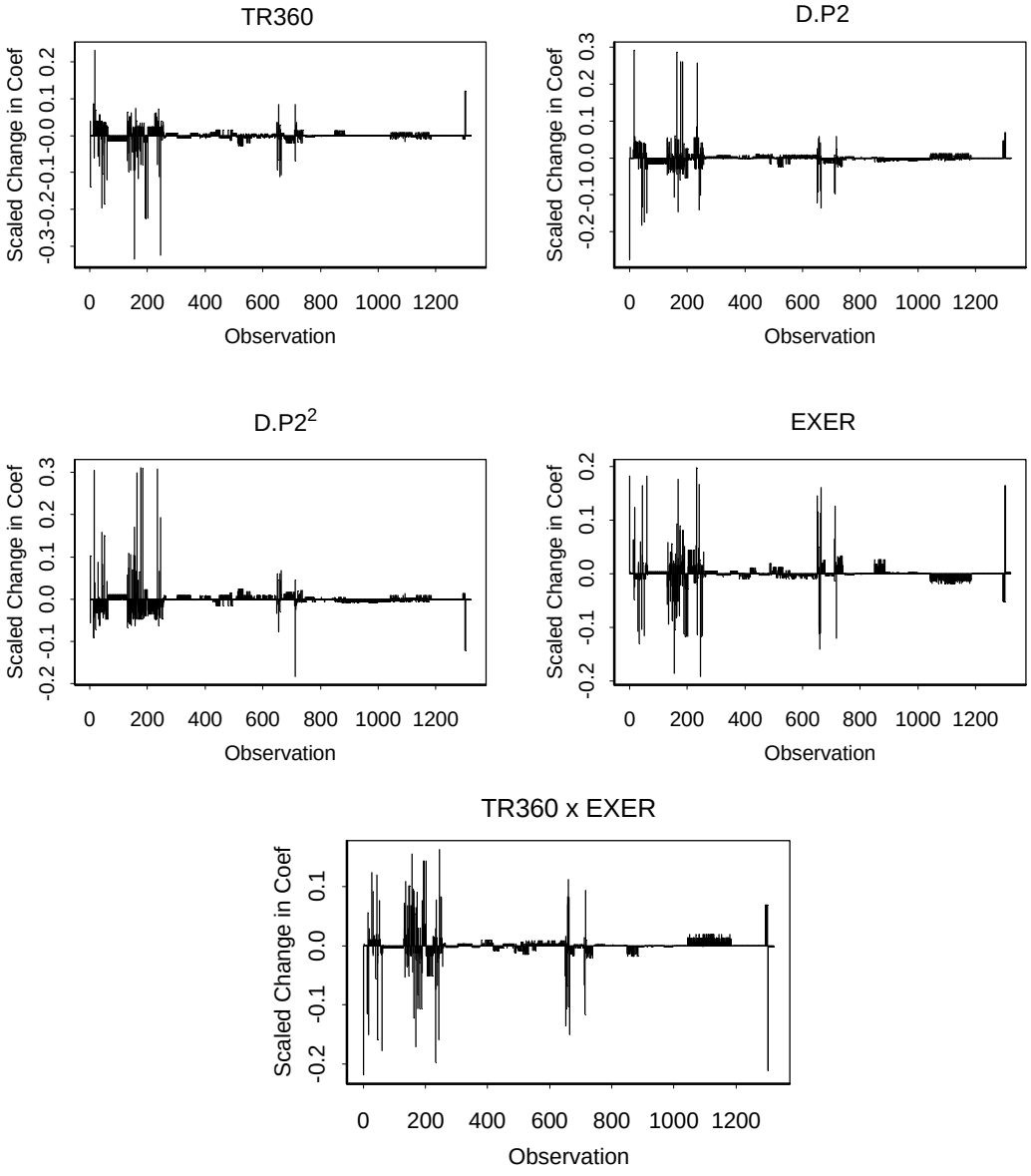

Figure 16: Influence for the covariates in Model 5, by observation.

Table 6: Weighted Partial Maximum Likelihood Estimates for Stratified Cox Model

	Model 6 (stratified on P2)
–2 Log LH	1447.38
AIC	1457.38
Parameter Estimates	
b_1 (TR360)	2.287 (0.409)
b_2 (P2$-\overline{\text{P2}}$)	–0.353 (0.124)
b_3(P2$-\overline{\text{P2}}$)^2	–0.343 (0.081)
b_4 (EXER)	–1.916 (1.088)
b_5 (TR360:EXER)	1.434 (0.577)

To assess the effect of the weighting on influence, in Figure 17 we show the dfbeta residuals for Model 6. For each covariate, we have plotted the observation by the approximate scaled change in the coefficient after removing the observation from the model. Weighting appears to have removed much of the influence of the early observations. However, some later observations now have dfbeta values greater than 30% of a standard error for P2, which corresponds to a change of only about 0.05 in the negative direction. One of these values corresponds to the only event in the fourth stratum of P2 values that is greater than 7.8.

Unfortunately, although the weighted model appears to fit well and removes some of the high influence values, it is unclear how to estimate survival with a weighted model for a given set of covariates. One option is to modify the weighted cumulative hazard function for known *case* weights given in Therneau (1999, p. 35). If we start with Breslow's estimate in (4.4), this means that we estimate the weighted cumulative hazard function within stratum *k* as

$$\hat{H}_{0_k}(t\,|\,\mathbf{x}_{(g)}) = \sum_{t_{i_k} \le t} \frac{w_{i_k} d_{i_k}}{W(t_{i_k}; \hat{b})} \tag{5.4}$$

where w_{i_k} is the appropriate weight at the i_kth event time, and $W(t_{i_k}; \hat{b}) = \sum_{j \in R(t_{i_k})} \exp(\mathbf{x}_j^T \hat{b})$. Note that the denominator does *not* involve the weights. This is an important distinction between a weighted likelihood and a Cox model with known case weights (Therneau, 1999). One complication with using (5.4) is that the computation of standard errors must involve the variance of the weights as they are estimated. Thus, we leave this topic for a later paper.

If we naively substitute the parameter estimates from Model 5 with the parameter estimates for Model 6 and then compute estimated survival as in previous sections, we get the survival estimates in Figures 18 and 19, which could be compared to Figures 12 and 13, respectively. Notice that the survival probabilities are predicted to be greater for 6.0 < P2 $\le$ 7.8 than for P2 $\le$ 4.3 at most hours, which is the opposite order to that seen in Figures 12 and 13. However, the confidence intervals cannot be computed so readily for this type of prediction from a weighted model.

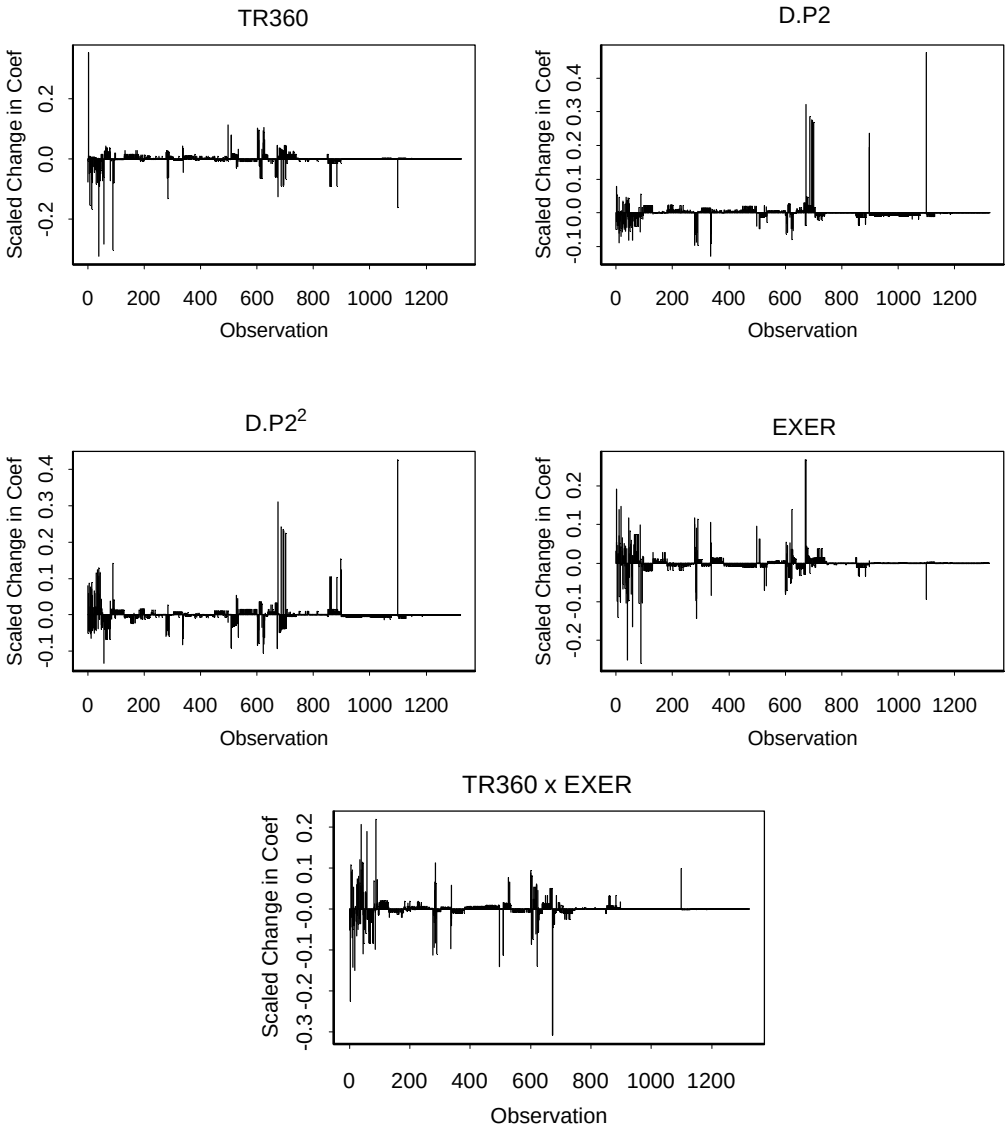

Figure 17: Influence for the covariates in Model 6, by observation.

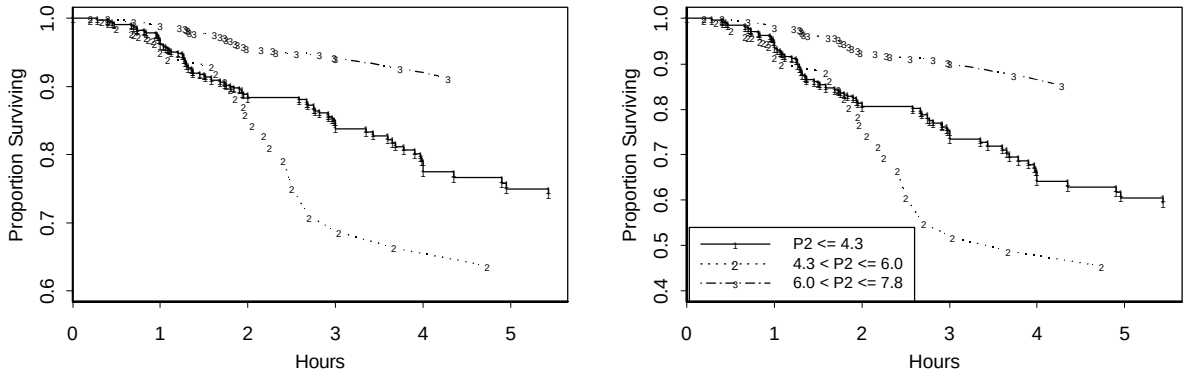

Figure 18: Expected survival for hypothetical individuals who exercised at altitude with (left) TR360 = 1.60 and (right) TR360 = 1.75.

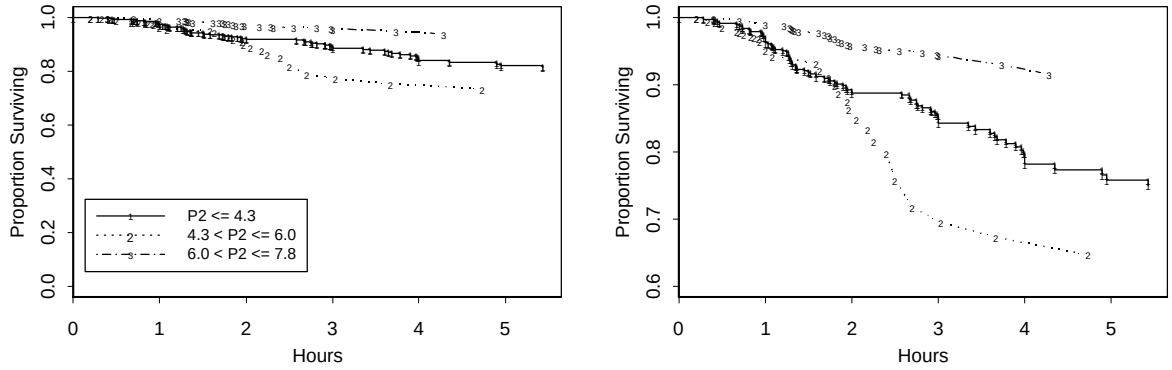

Figure 19: Expected survival for hypothetical individuals who did not exercise at altitude with (left) TR360 = 1.60 and (right) TR360 = 1.75.

6. Interpretation of the Stratified Cox Model

As we mentioned in Section 2, the exponents of the MPLEs in the Cox model with covariates are estimates of the relative risk of each covariate as compared to a baseline. For a stratified model, these estimates apply to all strata. Model 5 (Table 4) says that for exercisers, for each one-unit increase in TR360, when P2 is held constant within a stratum, the hazard increases over 73-fold. For non-exercisers, the DCS risk increases only over 13-fold for each one-unit increase in TR360. The modeled effect of P2 is always nonpositive but quadratic. Without considering stratification, estimates say that as P2 rises to its mean of 6.20, with TR360 kept constant, the effect on the hazard of DCS symptoms increases. As P2 increases beyond its mean, the effect on the hazard decreases. But within a stratum, the values of P2 are limited to those values the stratum represents. Note that for TR360 to be 'kept constant' while P2 varies, PN2360, the partial pressure of nitrogen must vary, too, because TR360 is a ratio of PN2360 to P2.

The modeled quadratic effect of P2 contrasts with the linear effect of P2 found by Chhikara et al. (1998). To check the authenticity of the quadratic effect on hazard, we used the methodology of Gray (1996) to compute a nonparametric estimate of the hazard based on covariates. Briefly, in Gray's method, the data are binned based on covariate groups (by specified quantiles) and time intervals, and a smoothed (LOWESS) estimate of the hazard is formed. Gray's method is implemented in the S-PLUS function *hazcov*. Figure 20 gives the hazard by time and P2. At any time point as P2 increases, the hazard rate appears quadratic, although not necessarily in the same way as the Cox

model estimates indicated. However, the Cox model estimates are independent of time and only describe an effect on a baseline hazard that changes over time.

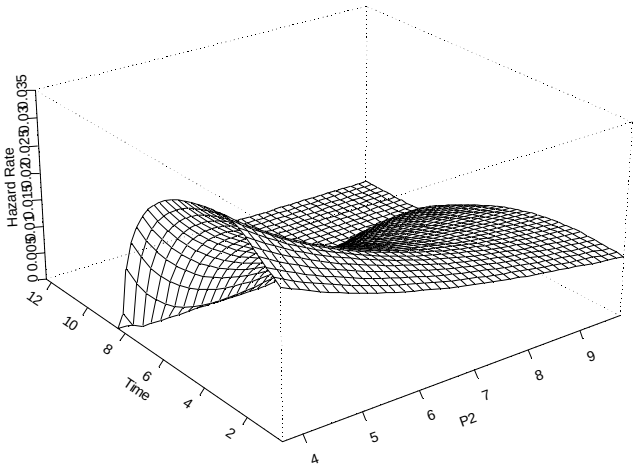

Figure 20: Hazard estimate by time and P2 (using Gray's *hazcov* function).

To get estimates of absolute risk instead of relative risk, we can use the estimated cumulative or integrated hazard function. But, this is just minus the log of the expected survival that was computed in Figures 12 and 13.

7. Discussion On the Use of Frailty Models for DCS Data

It is clear that a model stratified on quartiles of P2 suffers some lack of fit. In Section 5.5, we showed that some individuals are not fit well by the model, and some risk groups are not as well-described as others. There may also be unmeasured prognostic factors. These situations suggest that the addition of subject-specific frailty terms into the partial likelihood may help the fit of the stratified Cox model. A frailty is an unobserved continuous random variable that describes excess risk or 'frailty' for distinct groups, such as families or even single individuals, in addition to that described by measured covariates (Therneau et al., 2000). Thus, frailties are like unobserved covariates. Individuals with greater frailty are expected to experience the event earlier than those with lower frailties.

A Cox model with subject-specific frailties is a special case of a shared frailty model (Hougaard, 2000). The term *shared* comes from the use of such models to account for dependence among certain observations. In our case, we would have a single frailty per record. Thus, it is more appropriate to consider frailties as picking up excess variation not modeled by measured covariates. As the covariates are the same for each group of tested individuals instead of being unique to an individual, there may be some subject-specific variability in DCS times that is not already modeled.

In a frailty model, the hazard conditional on the frailty is

$$l(t;\mathbf{X},v) = l_0(t)v\ \exp(\mathbf{X}b) \tag{5.5}$$

where v is a positive random variable called a frailty, and is usually rewritten in the form $v = \exp(w)$ so that (5.5) becomes, in vector notation,

$$l(t;\mathbf{X},v) = l_0(t)\exp(\mathbf{X}b + \mathbf{Z}w) \tag{5.6}$$

with **Z** a vector of indicators that selects the appropriate w term for each subject. For many frailty distributions, proportional hazards are only assumed in the conditional hazard. The marginal hazard formulation resulting from integrating the associated survival distribution of (5.6) with respect to the frailty distribution does not usually reflect proportional hazards (Hougaard, 2000).

Using frailties in Cox models is quite common. In fact, it is believed by some (Hougaard, 1995) that all models should contain frailties. Henderson and Oman (1999) show that when frailty is present but ignored in a Cox model, the regression coefficients are biased towards zero. However, when censoring is present, the bias is reduced. In our data set, we have over 85% right-censoring. Thus, it is of interest to see if fitting frailties makes any difference in the coefficients. The addition of frailties to the Cox model is very similar to the addition of random effects. Previous work on the use of random effects in DCS research comes from Thompson et al. (2002) and Thompson and Chhikara (2001).

To use frailties in our model, we must specify a distribution for the frailties; that is, the distribution from which the frailties are assumed a random sample. There are many conventional choices for the frailty distribution in the literature. Hougaard (1995) reviews many of these choices. We considered only two choices: a gamma distribution and a lognormal distribution. Both of these frailty distributions allow us to estimate parameters by maximizing a penalized partial log likelihood with penalty function equal to the log likelihood for a random sample of w 's from the appropriate distribution (Therneau and Grambsch, 2000). This is conveniently implemented in the S-PLUS function *coxph.penal*. The parameters to be estimated are the coefficients in the ordinary Cox partial likelihood, plus any unknown parameters in the frailty distribution. The frailties themselves can also be estimated, if desired.

We fit several frailty models to the DCS data. We included stratification so that the conditional hazard function in (5.6) has a different baseline hazard per stratum. We fit two different frailty distributions, the gamma and lognormal, but we only give results for the gamma distribution, as the conclusions were similar between the two distributions. The gamma distribution allows a greater chance for some frailties to be near zero than does the lognormal distribution (Therneau and Grambsch, 2000). Because $v = \exp(w)$ is distributed gamma, w is distributed as log-gamma. It is the distribution of the log of v that is used in the penalty function. For purposes of identification, the mean of v is fixed at one. This leaves one unknown parameter—say, q —to describe the variance of w. The variance term can be estimated along with the ordinary Cox model parameters.

Table 7 gives results from the addition of frailty terms to Model 5. Estimates of the exponents of the coefficients b are the relative risks for any given subject. The log likelihood given for the gamma frailty model is the log partial unpenalized likelihood integrated with respect to the frailty distribution. A likelihood ratio test (LRT) that the frailty variance exceeds zero is given by twice the difference between this integrated log likelihood and the log likelihood of a model without frailties (Model 5). This statistic has an approximate chi-squared distribution with one degree of freedom (Therneau and Grambsch, 2000).

Estimation of the frailty variance q is done in an outer loop of a Newton-Raphson algorithm for estimating b and w. Assuming a fixed q, $\hat{b}$ and $\hat{w}$ are found by a Newton-Raphson algorithm. Then, $\hat{q}$ is found by maximizing the profile likelihood with b and w profiled out, and $\hat{b}$ and $\hat{w}$ are then re-estimated. Full details of the estimation procedures can be found in Therneau and Grambsch (2000) and Therneau et al. (2000).

Table 7: Partial Maximum Likelihood Estimates for Stratified Cox Model with Frailties

	Model 7 (Gamma frailty)
−2 Integrated LogLH	1732.00
LRT of frailty variance	0.10 (df = 1)
Parameter Estimates	
b_1 (TR360)	2.629 (0.526)
b_2 (P2$-\overline{\text{P2}}$)	0.149 (0.350)
b_3(P2$-\overline{\text{P2}}$)^2	−0.324 (0.108)
b_4 (EXER)	−2.548 (1.049)
b_5 (TR360:EXER)	1.781 (0.557)
q Frailty Variance	0.164 (0, 2.11)

Table 7 gives estimates that are very similar to those from Model 5. The standard errors for the coefficient estimates in Table 7 come from the inverse of the second derivative matrix of the penalized log likelihood. Because these estimates are computed assuming that the parameter q is fixed, they are underestimated. The standard errors can be corrected using the bootstrap procedure. The bootstrap procedure would have to be appropriate for censoring and for stratification (Davison and Hinkley, 1997). In addition, the standard error for q is not provided from standard statistical software. The bootstrap again can be used to obtain it instead. However, one suggestion by Therneau and Grambsch (2000) for the gamma frailty model is to compute a confidence interval for q using the profile likelihood. The profile likelihood confidence interval finds all values of q for which the LRT statistic (using the unpenalized likelihood) exceeds 3.84, the chi-squared 95th percentile for one degree of freedom. The interval is not usually symmetric about the point estimate. A profile likelihood-based confidence interval for q is (0, 2.11).

Based on the LRT and the confidence interval, we conclude that the estimated frailty variance does not significantly differ from zero. Thus, we can probably conclude that the measured covariates account for most of the variance in DCS times. Also, the coefficient estimates in Table 7 are very close to those in Table 5. It is interesting to note that for models that exclude certain terms, however, the frailty variance significantly differs from zero, implying that those terms are needed in the model. For example, if we use subject-specific frailty terms in the initial Cox model with three covariates and no interaction (Model 1), we get a highly significant frailty variance. The LRT statistic is 29.3 on df = 1, giving a p-value less than 0.001. Including the interaction between TR360 and EXER gives an LRT statistic of 4.08, giving a p-value of 0.04. The model fit by Chhikara et al. (1998) included the covariates EXER, P2, and PN2360. With gamma frailties added to this model, the frailty variance is estimated at 1.48, and the LRT is 5.92 giving a p-value of 0.02. As one major role of frailties is to pick up an excess variation that is not already modeled by terms in the model, we can conclude that previously considered models are likely inadequate for the DCS data.

8. Model Validation

In this section, we discuss model validation including predictive accuracy and calibration of predictions of a model applied to a future data set. We apply recently proposed measures of validation in the literature to the Cox model stratified on P2 quartiles, as well as to the other models, to compare them. In previous sections, we used AIC to compare the fit of various Cox models. AIC may be considered a measure of relative fit of a model to the existing

data. Here we compare the fit of several models in terms of how well they are expected to predict or explain survival on a future data set that is similar to the existing data set.

8.1. Predictive Accuracy of Cox Models

The predictive accuracy or predictive value of a statistical model measures its fit to future data that are similar to the data that were used to fit the original model. If a validation data set is available, the fitted model can be applied to this data set, and discrepancies in the predictions calculated, to obtain a measure of predictive value of the model. If a validation data set is not available, leave-one-out cross validation is an option for obtaining a measure of predictive value.

Verweij and Van Houwelingen (1993) describe a measure of predictive value from the fit of a Cox proportional hazards model. Predictions from the Cox model are made using the prognostic index (*PI*) $\mathbf{x}_i'\hat{b}$. First, a *PI* is obtained for each individual *i*, based on a model without that observation, $\mathbf{x}_i'\hat{b}_{(-i)}$. Then a Cox regression is performed on the complete data set with $\mathbf{x}_i'\hat{b}_{(-i)}$ as the only covariate. The partial log likelihood from this model is denoted by $l^*(c)$, where *c* is the regression coefficient for the *PI*. If $l(c)$ denotes the partial log likelihood from the fit of the original data to the *PI* $\mathbf{x}_i'\hat{b}$, then $c = 1$ maximizes $l(c)$. Thus, $l(1)$ is considered a measure of fit to the data from which the model was derived. Similarly, $l^*(1)$ measures the fit to future data. Thus, $l^*(1)$ is a measure of the predictive value of the model. In addition, Verweij and Van Houwelingen note that the coefficient estimate $\hat{c}$ that maximizes $l^*(c)$ is a shrinkage factor that can be used to estimate the amount by which regression coefficients are overestimated or exaggerated in the original Cox regression. A value of $\hat{c}$ close to 1.0 implies little overestimation. Applying this shrinkage factor to the linear combinations, $\mathbf{x}_i'\hat{b}$ will give adjusted survival estimates. In this way, predictions can be improved by shrinkage.

The values of $l^*(1)$ and $\hat{c}$ for Models 2 through 5 are given in Table 8. Standard errors for $\hat{c}$ are given in parentheses. Recall that Models 4 and 5 are stratified on EXER and P2, respectively. The stratification was also included in the calculation of $l^*(1)$. So, to interpret the values in Table 8 correctly, the future data set would have to be stratified on these variables as well. Thus, the same reasons for stratifying on EXER and P2 (namely, evidence of non-PH) must be true of the new data set. Furthermore, the cross-validation did not recompute quartiles of P2 for each fit because only one observation was left out each time. Thus, the same cut-points for P2 are assumed to apply to the new data set. This is reasonable because it is assumed that the new data will have similar values of the covariates as the original data.

Table 8: Predictive Accuracy of Models

	Model 2	**Model 3**	**Model 4**	**Model 5**
$l^*(1)$	-1044.542	-1037.02	-964.42	-872.34
$\hat{c}$	0.956 (0.06)	0.952 (0.07)	0.934 (0.07)	0.918 (0.09)

According to Table 8, Model 5 has the highest measure of predictive value but the lowest shrinkage factor. However, the standard errors on the shrinkage factors are large enough to indicate that any 'true' difference among them is very slight. The order in the $\hat{c}$ estimates reflects model complexity. The most complex model is Model 5 because it has four strata and five coefficients, and the least complex model is Model 2 because it has four coefficients. The most complex model requires the greatest amount of shrinkage when applied to a future data set. We use the shrinkage factors next to calibrate the model predictions.

8.2. Model Calibration of Survival Predictions

To calibrate survival predictions, we again use the methodology of Verweij and Van Houwelingen (1993). Verweij and Van Houwelingen improve individual survival predictions from a Cox model by adjusting the *PI* by the shrinkage factor, $\hat{c}$, above. Verweij and Van Houwelingen's adjusted prognostic index (*API*) is then

$$API = \bar{X}\hat{b} + \hat{c}(X - \bar{X})\hat{b} \tag{5.7}$$

Note that when $0 \le \hat{c} < 1$, (5.7) is pulled toward the mean estimate $\bar{X}\hat{b}$, and when c = 1, (1.17) equals the ordinary *PI*. Estimated expected survival curves that use the *API* in place of the *PI* are corrected for overestimation caused by overfitting because they are pulled in closer to the survival estimate evaluated at the mean covariate vector.

To correct for overfitting in the survival curves in Section 5.4, we use the *API* in (5.7) computed using the MPLEs from Model 5 in Table 4. This leads to the adjusted prediction model

$$S(t \mid \mathbf{x}) = S_0(t)^{\exp(API)} \tag{5.8}$$

where $S_0(t)$ is estimated using the Flemington-Harrington estimator mentioned in Section 5.4. It should be noted that the shrinkage factor is an estimate with an associated uncertainty. Thus, we can still expect some deviation if we were to apply this model to a new data set. Any confidence interval on (5.8) must account for the uncertainty in $\hat{c}$.

Figures 21 and 22 are the analogues of Figures 12 and 13, with the *API* substituted for the *PI*. As expected, the differences among the curves for each level of P2 are less pronounced than they are in Figures 12 and 13. For both TR360 values, the estimated expected survival prior to about 1.5 hours is roughly the same no matter the level of P2. After 1.5 hours, for P2 between 6.0 and 7.8 psia, the estimated survival probability decreases more rapidly, but the probabilities for P2 $\le$ 4.3 and 6.0 < P2 $\le$ 7.8 remain roughly equal because the curve for P2 $\le$ 4.3 was pulled closer to that of 6.0 < P2 $\le$ 7.8, which covers the mean P2.

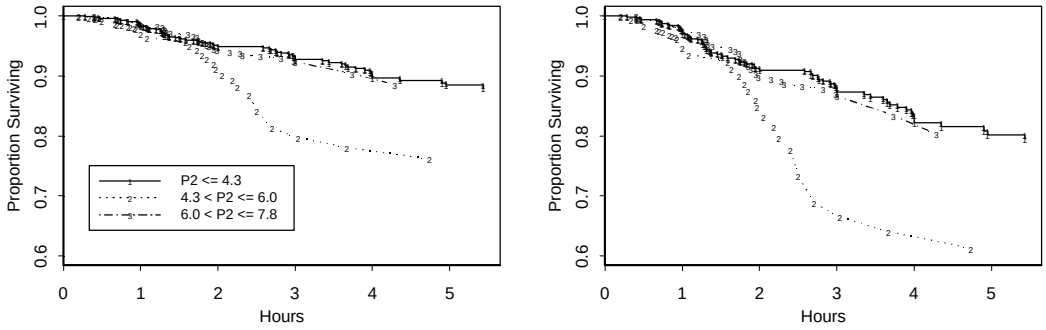

Figure 21: Expected survival for hypothetical individuals who exercised at altitude with (left) TR360 = 1.60 and (right) TR360 = 1.75.

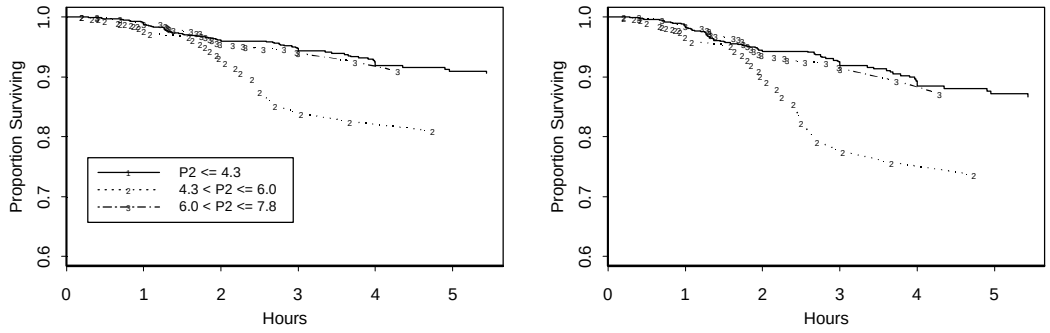

Figure 22: Expected survival for hypothetical individuals who did not exercise at altitude with (left) TR360 = 1.60 and (right) TR360 = 1.75.

9. Concluding Remarks

A Cox model stratified on quartiles of the final ambient pressure at altitude (P2) appears to fit the data adequately. However, some improvement may be made by a weighting procedure (Model 6). Since the addition of frailties did not appear to improve fit and the estimated frailty variance was not significantly greater than zero, we do not believe random effects are necessary to account for unmodeled covariates or there is substantial variability in the frailties of the subjects. However, this may be due to the nature of the data collection and not to any indication that DCS is not a highly personal occurrence. Subjects were tested in groups, and measured explanatory variables were applied to the whole group of individuals (although many groups were fairly small with two or three individuals). Had the grouping information been available, a shared frailty model could have accounted for some of the variability among groups of tested individuals.

A measure of predictive accuracy indicates that Model 5 has better predictive ability than the other models we considered. However, shrinkage of the prognostic index $\mathbf{x}'\hat{b}$ is expected to be higher for Model 5 due to its being the most complex of the models. Predicted survival curves using adjusted prognostic indices show a dampening of the effects of covariates on survival.

We have tried to adequately model DCS occurrence, but there still remains quite a bit of room for improvement. Various assessments of the proportion of explained variation in the data accounted for by the fitted Cox model (e.g., Schemper, 1990, 1992; Harrell, 1998) show that only about 25-40% of the variation is accounted for by any of the models we consider, and this is only on the data set at hand, not on future data sets.

Although the Cox model is very popular for survival data, it is not the only flexible model available. The Cox model with time-fixed covariates assumes a multiplicative effect of covariates on the baseline hazard (except if covariates enter through stratification). Alternatively, Aalen's (1980) additive hazard model models the hazard as an *additive* combination of covariate terms, where the coefficients in the linear combination may dependent on time, allowing the covariate effects to vary over time. Thus, covariates have an additive effect on the baseline hazard. This model measures additional excess risk due to the effects of a covariate in absolute terms instead of relative terms (Klein and Moeschberger, 1997).

In addition, frailty distributions can be nonparametric instead of having a form that is dependent on only a few parameters (Ibrahim et al., 2001). To achieve flexibility along with structure, we can use a scale mixture of Gaussian distributions for the distribution of w . This type of distribution might be used to check for outlying frailties as observations with particularly low scale estimates (Wakefield et al., 1994).

Appendix – Arjas Plots

Arjas plots (Arjas, 1988) can be used to graphically check the PH assumption for a given fitted Cox model, for each covariate in the model.

To check the PH assumption for a given covariate, say X_g, first, we fit a Cox model using all covariates except the gth. Let $\hat{b}$ be the MPLE from this fit. We then group the values of X_g into K categories and, for each failure time, t_{i_k}, in the kth category, we compute the expected cumulative number of failures in the kth category at that time as

$$E_k(t_{i_k}) = \sum_{j_k} \hat{H}_0(\min(t_{i_k}, T_{j_k}) | \mathbf{x}^*_{j_k}) \exp(\mathbf{x}^*_{j_k} \hat{b}) \tag{5.9}$$

where

$\hat{H}_0(\min(t_{i_k}, T_{j_k}) | \mathbf{x}^*_{j_k})$ is the estimated cumulative baseline hazard in **(4.4)**

T_{j_k} is the recorded failure time for the jth subject in the kth category

and $\mathbf{x}^*_{j_k}$ includes all covariates except X_g for the jth subject in the kth category

The observed cumulative number of failures that have occurred in the kth category up to time t_{i_k} is

$$N_k(t_{i_k}) = \sum_{j_k} d_{j_k} I(T_{j_k} \le t_{i_k}) \tag{5.10}$$

where d_{j_k} = 1 if the jth subject in the kth category has an uncensored recorded time, and is zero otherwise.

To create the Arjas plot, we plot $E_k(t_{i_k})$ by $N_k(t_{i_k})$ to compare observed and expected cumulative failures at time t_{i_k} for the kth category. Klein and Moeschberger (1997) give some guidelines for its interpretation. If the covariate does not belong in the model, then $N_k(t_{i_k}) - E_k(t_{i_k})$ is a zero-mean martingale, and a plot of $N_k(t_{i_k})$ by $E_k(t_{i_k})$ should be close to a 45-degree line through the origin. If the covariate belongs in the model and the correct model for the hazard is $h(t | X_g = k, \mathbf{X}^*) = h_o(t) \exp(g_k) \exp(b^T \mathbf{X}^*)$, then the Arjas plot will give graphs for each category that are approximately linear, but with slopes differing from one. If the omitted covariate X_g has a non-PH effect on the hazard rate, then the graphs will differ nonlinearly from the 45-degree line.

References

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes, *Lecture Notes in Statistics*, 2, 1-25.

Andersen, P. (1982). Testing goodness-of-fit of Cox's regression and life model, *Biometrics* 38, 67-77.

Arjas, E. (1988). A graphical method for assessing goodness-of-fit in Cox's proportional hazards model. *Journal of the American Statistical Association*, 83, 204-212.

Chhikara, R., Koti, K., and Spears, F. (1998). Cox Regression Modeling and Analysis of Decompression Sickness Data. ISSO Annual Report for 2000, University of Houston, TX.

Collet, D. (1994). *Modelling Survival Data in Medical Research*., London: Chapman & Hall.

Conkin, J., Bedahl, S., and Van Liew, H. (1992). A computerized databank of decompression sickness incidence in altitude chambers, *Aviation, Space, and Environmental Medicine*, 72, 202-214.

Conkin, J., Kumar, K., Powell, M., Foster, P., and Waligora, J. (1996). A probabilistic model of hypobaric decompression sickness based on 66 chamber tests, *Aviation, Space, and Environmental Medicine*, 67, 1-8.

Conkin, J., Powell, M., Foster, P., and Waligora, J. (1998). Information about venous gas emboli improves prediction of hypobaric decompression sickness, *Aviation, Space, and Environmental Medicine*, 69, 8-16.

Conkin, J. (2001). Evidence-based approach to the analysis of serious decompression sickness with application to EVA astronauts. NASA Technical Publication 2001-210196, Houston: Johnson Space Center, January 2001.

Cox, D. (1972). Regression models and life tables, *Journal of the Royal Statistical Society B*, 34, 187-202.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application*, Cambridge University Press.

English, T. (2000). *Goodness of fit test for the Cox proportional hazards model*, Unpublished Master's Thesis, University of Houston-Clear Lake, Houston, TX.

Fleming, T. and Harrington, D. (1984). Nonparametric estimation of the survival distribution in censored data, *Communications in Statistics: Theory and Methods*, 13, 2469-2486.

Foster, P., Conkin, J., Powell, M., Waligora, J., and Chhikara, R. (1998). Role of metabolic gases in bubble formation during hypobaric exposures, *Journal of Applied Physiology*, 1088-1095.

Grambsch, P. and Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81, 515-526.

Gray, R. (1996). Hazard rate regression using ordinary nonparametric regression smoothers, *Journal of Computational and Graphical Statistics*, 5, 190-207.

Harrell, F. (1998). *Predicting Outcomes: Applied Survival Analysis and Logistic Regression*, University of Virginia, Charlottesville.

Henderson, R. and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis, *Journal of the Royal Statistical Society Series B*, 61, 367-379.

Hess, K., Serachitopol, D., and Brown, B. (1999). Hazard function estimators: A simulation study, *Statistics in Medicine*, 18, 3075-3088.

Hosmer, D. and Lemeshow, S. (1998). *Applied Survival Analysis*, New York: Wiley.

Hougaard, P. (1995). Frailty models for survival data, *Lifetime Data Analysis*, 1, 255-273.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, New York: Springer-Verlag.

Ibrahim, J., Chen, M., and Sinha, D. (2001). *Bayesian Survival Analysis*, New York: Springer-Verlag.

Kannan, N. and Raychaudhuri, A. (1998). *Survival Models for Predicting Altitude Decompression Sickness*. USAF Technical Report AL/CF-TR-1997-0030, Brooks AFB, San Antonio, Texas.

Klein, J. and Moeschberger, M. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer-Verlag.

Kumar, K. and Powell, M. (1994). Survivorship models for estimating the risk of decompression sickness, *Aviation, Space, and Environmental Medicine*, 65, 661-665.

Lawless, J. (1982). *Statistical Models and Methods for Lifetime Data*, New York: Wiley

Lin, D. and Wei, L. (1989). The robust inference for the Cox proportional hazards model, *Journal of the American Statistical Association*, 84, 1074-1078.

May, S. and Hosmer, D. (1998). A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model, *Lifetime Data Analysis*, 4, 109-120.

Nardi, A. and Schemper, M. (1999). New residuals for Cox regression and their application to outlier screening, *Biometrics*, 55, 523-529.

Parzen, M. and Lipsitz, S. (1999). A global goodness-of-fit statistic for Cox regression models, *Biometrics*, 55, 580-584.

Schemper, M. (1990). The explained variation in proportional hazards regression, *Biometrika*, 77, 216-218 (correction in 1994, 81, 631).

Schemper, M. (1992). Cox analysis of survival data with non-proportional hazard functions, *The Statistician*, 41, 455-465.

Schemper, M. (1992). Further results on the explained variation in proportional hazards regression, *Biometrika*, 79, 202-204.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model, *Biometrika*, 67, 145-153.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika*, 69, 239-241.

S-PLUS (2001). *Guide to Statistics, Vol. 2*, Insightful Corporation, Cambridge, MA.

Therneau, T. (1999). *A package for survival analysis in S.* Technical Report Series No. 53, Department of Health Science Research, Mayo Clinic, Rochester, MN.

Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*, New York: Springer-Verlag.

Therneau, T., Grambsch, P. and Pankrantz, S. (2000). *Penalized Survival Models and Frailty*. Technical Report Series No. 66, Department of Health Science Research, Mayo Clinic, Rochester, MN.

Thompson, L. and Chhikara, R. (2001). *Estimating the risk of serious DCS over the lifespan of the ISS,* ISSO Annual Report for 2001, University of Houston, TX.

Thompson, L., Conkin, J., Chhikara, R., and Powell, M. (2002). *Modeling Grade IV venous gas emboli using a limited failure population model with random effects.* NASA Technical Publication 2002-210781, Houston: Johnson Space Center, May 2002.

Valsecchi, M., Silvestri, D., and Sasieni, P. (1996). Evaluation of longer-term survival: Use of diagnostics and robust estimators with Cox's proportional hazards model, *Statistics in Medicine,* 15, 2763-2780.

Verweij, P. and Van Houwelingen, H. (1993). Cross-validation in survival analysis, *Statistics in Medicine,* 12, 2305-2314.

Wakefield, J. Smith, A. Racine-Poon, A., and Gelfand, A. (1994). Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler, *Applied Statistics,* 43, 201-221.

Wilcox, R. (2001). *Fundamentals of Modern Statistical Methods*, New York: Wiley.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 2003	NASA Technical Paper

4. TITLE AND SUBTITLE
Cox Proportional Hazards Models for Modeling the Time To Onset of Decompression Sickness in Hypobaric Environments

5. FUNDING NUMBERS

6. AUTHOR(S)
Laura A. Thompson, Raj S. Chhikara, Johnny Conkin

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lyndon B. Johnson Space Center
Houston, Texas 77058

8. PERFORMING ORGANIZATION REPORT NUMBERS
S-890

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
TP-2003-210791

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Available from the NASA Center for AeroSpace Information (CASI)
7121 Standard
Hanover, MD 21076-1320 Category: 52

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In this paper we fit Cox proportional hazards models to a subset of data from the Hypobaric Decompression Sickness Databank. The data bank contains records on the time to decompression sickness (DCS) and venous gas emboli (VGE) for over 130,000 person-exposures to high altitude in chamber tests. The subset we use contains 1,321 records, with 87% censoring, and has the most recent experimental tests on DCS made available from Johnson Space Center. We build on previous analyses of this data set by considering more expanded models and more detailed model assessments specific to the Cox model. Our model – which is stratified on the quartiles of the final ambient pressure at altitude – includes the final ambient pressure at altitude as a nonlinear continuous predictor, the computed tissue partial pressure of nitrogen at altitude, and whether exercise was done at altitude. We conduct various assessments of our model, many of which are recently developed in the statistical literature, and conclude where the model needs improvement. We consider the addition of frailties to the stratified Cox model, but found that no significant gain was attained above a model that does not include frailties. Finally, we validate some of the models that we fit.

14. SUBJECT TERMS	15. NUMBER OF PAGES	16. PRICE CODE
decompression sickness, extravehicular activity, Hypobaric Decompression Sickness Databank, Cox proportional hazards model, censoring, frailty model, model validation	52	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unlimited	Unlimited